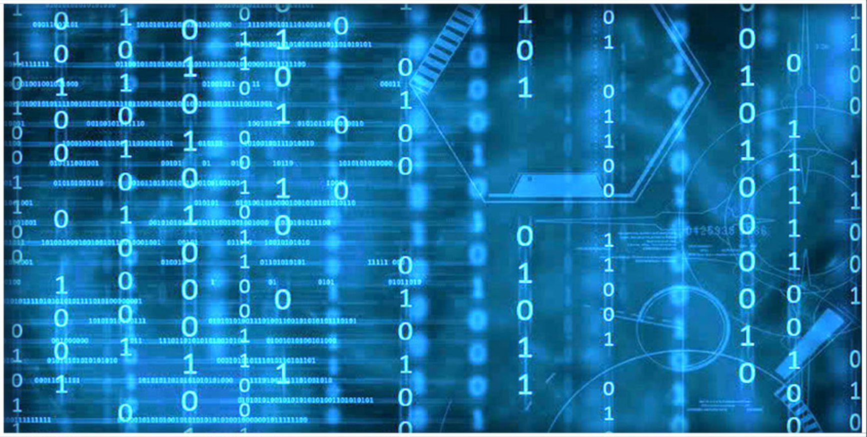
IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 11 Issue 3

March 2020



ISSN 2156-5570(Online)
ISSN 2158-107X(Print)

SAI

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

It may be difficult to imagine that almost half a century ago we used computers far less sophisticated than current home desktop computers to put a man on the moon. In that 50 year span, the field of computer science has exploded.

Computer science has opened new avenues for thought and experimentation. What began as a way to simplify the calculation process has given birth to technology once only imagined by the human mind. The ability to communicate and share ideas even though collaborators are half a world away and exploration of not just the stars above but the internal workings of the human genome are some of the ways that this field has moved at an exponential pace.

At the International Journal of Advanced Computer Science and Applications it is our mission to provide an outlet for quality research. We want to promote universal access and opportunities for the international scientific community to share and disseminate scientific and technical information.

We believe in spreading knowledge of computer science and its applications to all classes of audiences. That is why we deliver up-to-date, authoritative coverage and offer open access of all our articles. Our archives have served as a place to provoke philosophical, theoretical, and empirical ideas from some of the finest minds in the field.

We utilize the talents and experience of editor and reviewers working at Universities and Institutions from around the world. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations. Our high standards are maintained through a double blind review process.

We hope that this edition of IJACSA inspires and entices you to submit your own contributions in upcoming issues. Thank you for sharing wisdom.

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 11 Issue 3 March 2020
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)
©2013 The Science and Information (SAI) Organization

Editorial Board

Editor-in-Chief

Dr. Kohei Arai - Saga University
Domains of Research: Technology Trends, Computer Vision, Decision Making, Information Retrieval, Networking, Simulation

Associate Editors

Chao-Tung Yang
Department of Computer Science, Tunghai University, Taiwan
Domain of Research: Software Engineering and Quality, High Performance Computing, Parallel and Distributed Computing, Parallel Computing

Elena SCUTELNICU
"Dunarea de Jos" University of Galati, Romania
Domain of Research: e-Learning, e-Learning Tools, Simulation

Krassen Stefanov
Professor at Sofia University St. Kliment Ohridski, Bulgaria
Domains of Research: e-Learning, Agents and Multi-agent Systems, Artificial Intelligence, Big Data, Cloud Computing, Data Retrieval and Data Mining, Distributed Systems, e-Learning Organisational Issues, e-Learning Tools, Educational Systems Design, Human Computer Interaction, Internet Security, Knowledge Engineering and Mining, Knowledge Representation, Ontology Engineering, Social Computing, Web-based Learning Communities, Wireless/ Mobile Applications

Maria-Angeles Grado-Caffaro
Scientific Consultant, Italy
Domain of Research: Electronics, Sensing and Sensor Networks

Mohd Helmy Abd Wahab
Universiti Tun Hussein Onn Malaysia
Domain of Research: Intelligent Systems, Data Mining, Databases

T. V. Prasad
Lingaya's University, India
Domain of Research: Intelligent Systems, Bioinformatics, Image Processing, Knowledge Representation, Natural Language Processing, Robotics

CONTENTS

Apple Carving Algorithm to Approximate Traveling Salesman Problem from Compact Triangulation of Planar Point Sets

Marko Dodig[1], Milton Smith[2]
Industrial, Manufacturing and Systems Engineering
Texas Tech University
Lubbock, TX, USA

***Abstract*—We propose a modified version of the Convex Hull algorithm for approximating minimum-length Hamiltonian cycle (TSP) in planar point sets. Starting from a full compact triangulation of a point set, our heuristic "carves out" candidate triangles with the minimal Triangle Inequality Measure until all points lie on the outer perimeter of the remaining partial triangulation. The initial candidate list consists of triangles on the convex hull of a given planar point set; the list is updated as triangles are eliminated and new triangles are thereby exposed. We show that the time and space complexity of the "apple carving" algorithm are O(n^2) and O(n), respectively. We test our algorithm using a well-known problem subset and demonstrate that our proposed algorithm outperforms nearly all other TSP tour construction heuristics.**

Keywords*—TSP; heuristics; combinatorial optimization; computational geometry; compact triangulation

I. Introduction

In this article we examine the following tour-construction heuristic for the planar TSP: take a compact triangulation of the planar set and then find the minimum Hamiltonian cycle embedded in the triangulation by progressively removing triangles of minimal Triangle Inequality measure until n-2 triangles remain. We call this heuristic "apple carving" as this descriptor accurately describes the triangle removal process which is the basis of the algorithm. Possibility of using well-known triangulations such as Greedy and Delaunay to generate heuristic tours was already explored by Reinelt [1], Stewart [2], and Letchford and Pearson [3]. These authors looked at triangulations as presenting a "good" subset of edges and utilized well-established TSP solutions engines like CONCORDE to solve for TSP. Our research is different in that we (a) utilize newly introduced Greedy Compact Triangulation (GCT) proposed recently by Dodig and Smith [4], and (b) utilize a modification of Convex Hull Heuristic on GCT triangles to approximate TSP.

Our paper is organized as follows. First, we formally define the TSP and review the present state of its solution algorithms. Second, we introduce our approach. Third, we present our experimental methodology and review our experimental results. Finally, we highlight our conclusions and outline future research steps.

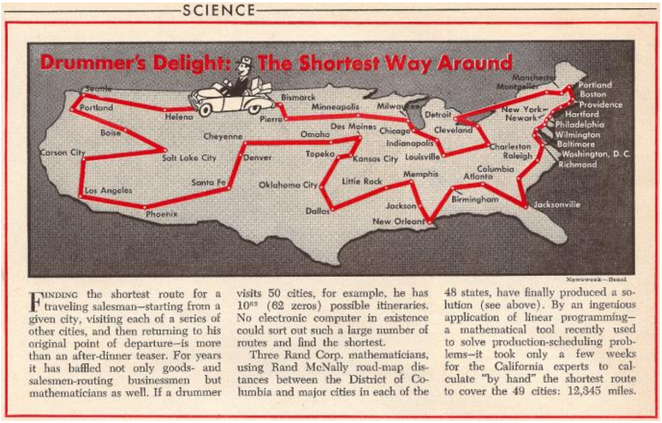
II. Literature Review

A. Traveling Salesman Problem

Traveling salesman problem (TSP) is perhaps the best-known and most-researched problem in combinatorial optimization. In its general form we are given a collection of cities and the distance to travel between each pair of them, and the problem then is to find the shortest route to visit each city and to return to the starting point [5]. TSP belongs to the class of NP-hard problems; in other words no polynomial-time algorithm exists that can solve the problem optimally in polynomial time, regardless of its complexity (i.e. the number of cities in the tour). The best result to date is a solution method, discovered in 1962, that runs in time proportional to $n^2 2^n$ [6]. TSP has been fascinating both researchers and general public for more than sixty years. In 1954, three researchers from Rand Corporation had solved a long-standing public challenge to find the shortest tour through 48 US state capitals and DC, shown in Fig. 1 [5].

In purely mathematical terms, TSP is the problem of finding a Hamiltonian tour (cycle) of minimum weight in a complete edge-weighted graph. In our research, we consider a symmetric TSP, or STSP, in that we assume that edge-costs are symmetric, or, equivalently, that the graph is undirected. A special case of the TSP is obtained when the vertices of the graph correspond to points in the Euclidean plane, and distance between any two points is equal to the Euclidean distance between the corresponding points. The Euclidean TSP is a special case of the metric TSP, in which the costs obey the triangle inequality. Metric TSP was found to be strongly NP-hard [7]. Related to, but distinct from, the Euclidean TSP is the planar graph TSP which is the focus of our research. This is the version of the TSP in which a planar graph G = (V, E) is given, with weights on the edges of E, and one seeks the minimum cost tour which uses only edges in E. Not only is this problem NP-hard, it is NP-hard even to test if a planar graph is Hamiltonian [7].

There is a multitude of planar TSP solution algorithms; few are exact algorithms, and many are heuristic algorithms. Since planar TSP is NP-hard, exact algorithms are exponential and heuristic algorithms are polynomial; selecting between exact or heuristic algorithms to solve for TSP presents a clear case of precision and time trade-off.



SCIENCE

Drummer's Delight: The Shortest Way Around

Newsweek—Bensi

FINDING the shortest route for a traveling salesman—starting from a given city, visiting each of a series of other cities, and then returning to his original point of departure—is more than an after-dinner teaser. For years it has baffled not only goods- and salesmen-routing businessmen but mathematicians as well. If a drummer visits 50 cities, for example, he has 10^{62} (62 zeros) possible itineraries. No electronic computer in existence could sort out such a large number of routes and find the shortest.

Three Rand Corp. mathematicians, using Rand McNally road-map distances between the District of Columbia and major cities in each of the 48 states, have finally produced a solution (see above). By an ingenious application of linear programming—a mathematical tool recently used to solve production-scheduling problems—it took only a few weeks for the California experts to calculate "by hand" the shortest route to cover the 49 cities: 12,345 miles.

Fig 1. Newsweek Coverage of 49-City Tour through United States [5].

B. Exact Algorithms

Branch-and-bound algorithm is an exact algorithm based on the IP formulation of TSP. This algorithm consists of two steps, (a) branching, which means splitting the problem into sub-problems, and (b) bounding, which means calculating lower and/or upper bounds for the objective function value of the sub-problem. The branching is performed in the following algorithm by separating the current subspace into two parts using the integrality requirement. Using the bounds, unpromising sub-problems can be eliminated. LP-relaxation of the problem is formed by relaxing integer requirements. In the algorithm, a list of sub-problems is maintained. A sub-problem is fathomed (totally solved) and removed from the list only when it has an integer solution that is best so far and becomes the new incumbent solution, or its optimum LP-solution objective is worse than the current incumbent value, or its LP-problem is infeasible.

Held-Karp algorithm is a dynamic programming algorithm utilizing graph theoretical representation of TSP. In a way, it is an intelligent brute force method in that it utilizes recursive formulation to find minimal distance paths between points. It was proposed independently by Bellman [6] and by Held and Karp [8]. This algorithm utilizes an optimization property of TSP in that every sub-path of a path of minimum distance is itself of minimum distance, which is easily proven by contradiction. The algorithm computes the solutions of all sub-problems, starting with the smallest, and looks up solutions already computed when requiring solutions for smaller problems. At the end, computing minimum distance tour means using the final equation to generate the initial node, and then repeating for all other nodes. Held-Karp is exhaustive, in that all sub-problems need to be solved; it has the time complexity of $O(2^n n^2)$ and the space complexity of $O(2^n n)$.

C. TSP Heuristics

In simplest terms, TSP heuristics can be divided into two distinct categories. Tour construction heuristics execute a sequence of operations until a valid tour is obtained, at which point the heuristics stop and report the constructed tour. Tour improvement heuristics start with a valid tour (an output of a tour construction heuristic, for example) and iteratively improve the tour cost, typically via local search, until some stopping criterion is reached [5]. Solution quality of tour improvement techniques far exceeds quality of solutions achieved by tour constructions [5].

Nearest Neighbor heuristic is perhaps the best-known tour construction heuristics [9]. It starts with a random city, adds the nearest non-visited city, and keep adding new non-visited cities in the same fashion until all cities are included. When all of the cities are included it returns to the initial city. It has the time and the space complexity of $O(n^2)$ and $O(n)$, respectively [10].

Greedy heuristic gradually constructs a tour by repeatedly selecting the shortest remaining edge and adding it to the tour as long as it does not create a cycle with less than n edges nor increase the degree of any node (city) to more than two [10]. Greedy heuristic has the time complexity of $O(n \times \log_2 n)$, which makes it more efficient than Nearest Neighbor [10]. The space complexity of Greedy matches that of Nearest Neighbor heuristic [10].

Cheapest Insertion heuristic starts with the shortest edge which becomes the initial sub-tour. Then it selects a city not in the current sub-tour, having the shortest distance to any one of the cities in the sub-tour. It finds an edge in the sub-tour such that the cost of inserting the selected city between the edge cities will be minimal, and keeps inserting shortest-distance remaining cities until none remain. Cheapest Insertion has the time complexity of $O(n^2 \times \log_2 n)$ and is more computationally intensive then Nearest Neighbor and Greedy [11].

Convex Hull heuristics starts by finding the convex hull of a point set and making it an initial sub-tour. For each remaining point it finds its cheapest insertion, adds the city with the least cost/increase ratio, and keeps repeating this process with remaining points until none remain. It is also more computationally intensive with the time complexity of $O(n^2 \times \log_2 n)$ [12].

Christofides heuristic builds a minimal spanning tree (MST) of the planar point set. It then creates a minimum-weight matching (MWM) on points having an odd degree, adds the MST together with the MWM, creates an Euler cycle from the combined graph, and finally traverses it taking shortcuts to avoid already included points. This heuristic has the best worst-case performance guarantee of all TSP heuristics as it never produces tours worse than 1.5 times the optimal [13]. On the other hand, it has the time complexity equal to $O(n^3)$ [13].

Match-Twice-and-Stitch heuristic [14] uses two sequential minimum-weight matchings to construct the cycles. The first matching returns the usual minimum-cost edge set with each point incident to exactly one matching edge. The second matching returns the minimum-cost edge set with each point incident to exactly one matching while ignoring the edges found in the first matching. The first phase results in multiple sub-tours. The second phase stitches the constructed cycles to form the TSP tour, with the exact (slow) and approximate (fast) patching procedure to join two cycles. A minimum spanning tree (MST) calculation determines a way to stitch all cycles into a tour. It is the best construction heuristics reported, with the different versions of the heuristic reporting average tour lengths between 4.8% (slowest) to 7.1% (fastest) over HK bound. It has the time complexity of $O(n^2)$ [14].

Tour improvement algorithm such as 2-opt removes two edges from the feasible tour and reconnects the two paths created if the new tour will be shorter. There is only one way to reconnect the two paths and still have a valid tour. It continues removing and reconnecting the tour until no 2-opt improvements can be found. Algorithm works the same for any path connecting k points, however the time performance severely lags starting at 5-opt. Its worst-case performance guarantee is known, as it is guaranteed to produce results not more than two times the optimal [10]. The main weakness of the 2-opt tour improvement heuristic is that it covers local improvements for pairs of 2 nodes only. This was subsequently addressed in newer k-opt algorithms, where k > 2, chief among them the Lin-Kernighan heuristic with the time complexity of O($n^{2.2}$) [10].

Solutions generated by TSP heuristics are typically compared to the Held-Karp (HK) lower bound. This lower bound is the solution to the LP relaxation of the IP formulation of the TSP, which can be found in polynomial time by using the Simplex method and a polynomial constraint-separation algorithm [15]. A HK lower bound averages about 0.8% below the optimal tour length [15]; however, its guaranteed lowest bound is only 2/3 of the optimal tour. Fig. 2 summarizes typical performance of the most-significant TSP heuristic algorithms. 2-opt, 3-opt, and Lin-Kernighan heuristics are the tour improvement heuristics, and all of the others are tour construction heuristics.

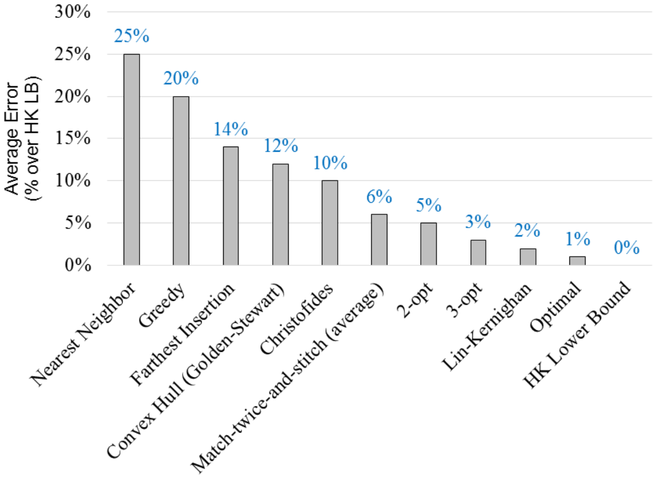


Fig 2. Typical Performance of Best-known Heuristics [10], [14].

III. Our Approach

A. Improved Greedy Compact Triangulation (iGCT)

iGCT of a planar point set S is created by GCT Algorithm [4]. This algorithm progressively inserts most-compact empty triangles into the triangulation not intersecting empty triangles in S previously inserted and achieves local optimality by performing weight-reducing edge flipping [4]. Compactness of an empty triangle Δ_T with area A(Δ_T) and perimeter P(Δ_T) in planar point set S is measured as follows [16]:

$$\mathrm{CI}(\Delta_T) = \frac{4\pi A(\Delta_T)}{[P(\Delta_T)]^2} \qquad (1)$$

Dodig and Smith showed that GCT approximates Minimum Weight Triangulation (MWT) in a variety of planar point set configurations, thereby making its edges compelling candidates for our proposed TSP heuristic [4]. MWT is defined as the full triangulation of a planar point set S having the lowest total edge length out of all full triangulations of a planar point set S. Dodig and Smith have also confirmed that the optimal TSP solution is frequently fully embedded in iGCT (61% of the time), and that the minimum perimeter polygon fully contained in iGCT is nearly optimal, or 0.36% longer than optimal. Fig. 3 shows full embeddedness of the optimal TSP tour in iGCT for *berlin52*, one of the TSPLIB problems for which the optimal TSP is known.

B. Apple Carving Algorithm

There are 2n - h - 2 triangles in both iGCT and MWT triangulations of a planar set S of n points, where h represents the number of points on the Convex Hull of S, or CH(S) [16]. We know that the perimeter length of CH(S) is less than the perimeter length of TSP polygon for this planar point set due to Isoperimetric Inequality principle. Following Steiner proof of Isoperimetric Inequality, we can “carve out” from CH(S) a triangle on the perimeter of full triangulation with the lowest Triangle Inequality Factor and have high degree of confidence that minimum perimeter polygon is still fully contained in the resulting partial triangulation. We can continue carving out eligible triangles with the lowest Triangle Inequality Measure, until all points are at the perimeter of the partial triangulation. We give priority to removing triangles whose absolute Triangle Inequality, or TI, is not only lowest, but also “optimal”. “Optimal” TI on any point is defined as the lowest TI of all triangles containing this point. We consider this method to be the basis of the “apple carving” algorithm. In fact, this method is very similar to the Convex Hull heuristics, through Convex Hull Heuristics does not follow a pre-defined tour building roadmap such as the one provided by the compact triangulation [12]. “Apple carving” algorithm pseudocode is given in Fig. 4.

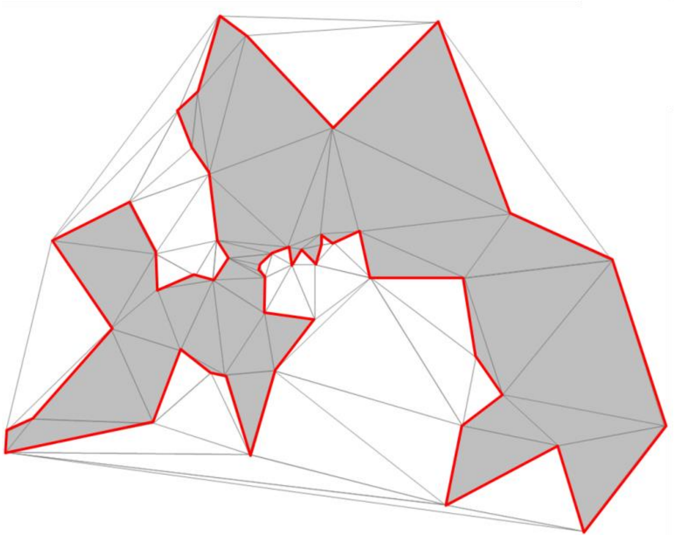


Fig 3. Optimal TSP (Shaded) Fully Contained in GCT for *berlin52* Problem [4].

```
INPUTS

    1. Planar point set S with n points; S has h points on CH(S).
    2. iGCT(S) with 2n – h – 2 triangles; each Triangle(a, b, c) and Edge(a, b) in iGCT satisfies a ≤ b ≤ c

BEGIN Apple Carving Algorithm

    1. Initialize variables
    2. Import point coordinates
    3. Initialize iGCT
        FOR each triangle i in iGCT
            SimpleTriangle(i) := Triangle(a, b, c)
            CountTriangles(a) +=1; CountTriangles(b)+=1; CountTriangles(c) +=1
            IF TIA (a, SimpleTriangle(i)) < min_TI (a) THEN
                min_TI(a) := TIA(a, SimpleTriangle(i))
            ENDIF
            CountEdges(a,b) += 1; CountEdges(a,c) += 1; CountEdges(b,c) += 1
            Apple ← SimpleTriangle(i)
        NEXT i
    4. Initialize Candidate List
        FOR each Edge(a, b)
            IF CountEdges(a, b) = 1 THEN
                CandidatesList ←  Edge(a, b)
                VisitedCities ← a, b
                TourLength += Distance(a, b)
            ENDIF
        NEXT
    5. Carve triangles from Polygon (Apple)
        change_recorded := 1
        WHILE VisitedCities < n AND change_recorded == 1
            change_recorded := 0
            Let k be the index of a triangle containing the candidate edge Edge(a, b) such that:
                a) CountTriangles(a) > 1 AND CountTriangles (b) > 1 AND CountTriangles(c) > 1,
                b) Min_TI(c) == TIA(c, SimpleTriangle(k))
                c) SimpleTriangle(k) == Triangle(a, b, c) with the min_TI(c) for all triangles satisfying a) and b)
            IF SimpleTriangle(k) doesn’t exist THEN
                Let k be the index of a triangle containing any candidate edge Edge(a,b) such that:
                    d) CountTriangles(a) > 1 AND CountTriangles(b)>1 AND CountTriangles(c) > 1,
                    e) SimpleTriangle(k) = Triangle(a, b, c) with the lowest TIR(c, SimpleTriangle(k)) for all triangles
                       satisfying d)
            ENDIF
            Apple → SimpleTriangle(k)
            CandidatesList ← Edge(a,c), Edge(b,c)
            CandidatesList → Edge(a,b)
            CountTriangles(a) -= 1; CountTriangles(a) -= 1; CountTriangles(a) -= 1
            VisitedCities ← c
            TourLength := TourLength - Distance(a, b) + Distance(a, c) + Distance(b, c)
            VisitedCities += 1; change_recorded := 1
        WHILE END
    6. Correct infeasibility conditions (if any)
        IF VisitedCities < n THEN
            FOR each point c NOT in VisitedList
                Let a and b be points in S such that
                    f) Edge(a,b) is in CandidatesList,
                    g) Triangle(a,b,c) has the lowest TIA(c, Triangle(a,b,c)) for any pair of points a and b satisfying f)
                VisitedCities ← c
                CandidatesList ← Edge(a,c), Edge(b,c)
                CandidatesList → Edge(a,b)
                TourLength = TourLength - Distance(a, b) + Distance(a, c) + Distance(b, c)
                VisitedCities += 1
            NEXT c
        ENDIF
    7. Record the polygon tour
        FOR each Edge(a, b) in CandidatesList
            Predecessor(b) := a
        NEXT

END Apple Carving Algorithm
```

Fig 4. Apple-Carving Algorithm Pseudocode.

C. Measure of Sub-optimality

We define $\varepsilon_{P''}^{A}(s)$ as the absolute deviation (from the optimal TSP) of the perimeter length of the polygon found via "apple carving" algorithm, and express it mathematically as follows:

$$\varepsilon_{P''}^{A}(s) = \frac{PL(P''(S))) - PL(TSP(S))}{PL(TSP(S))} \times 100\%, \forall S \ in \ R^2 \qquad (2)$$

Where S is a given point set, and P'' is the Hamiltonian cycle found by the "apple carving" algorithm.

D. Time Complexity

Theorem 1 The time complexity of the "apple-carving" algorithm is O(n^2).

Proof: Step 2 of the "apple carving" algorithm has the time complexity of O(n), since in this step we initialize arrays of n points. Step 3 of the "apple carving" algorithm has the time complexity of O(n), as we also know that there are O(n) triangles in a full triangulations of a planar point set S of n points [16]. Step 4 of the "apple carving" algorithm loops through no more than n candidate edges, and therefore has time complexity of O(n). Step 5 of the "apple carving" algorithm removes up to n – h triangles from iGCT. In each removal step, we evaluate up to 2n – h – 2 candidate triangles that can be removed. This guarantees time complexity of O(n^2) for Step 5. Step 6 of the "apple carving" algorithm has time complexity of O(n^2). We know this because there are not more than n points that need to be evaluated against up to n candidate edges/triangles. Finally, step 7 of the "apple carving" algorithm assigns predecessors for each of n points in S by looping through not more than n edges in the candidate lists, guaranteeing the time complexity of O(n).

This proves that the time complexity of the "apple carving" algorithm is 4O(n) + 2O(n^2) = O(n^2).

Theorem 2 The time complexity of the "apple-carving" algorithm and iGCT algorithm together is O(n^4).

Proof: Time complexity of the stand-alone "apple carving" algorithm is O(n^2). Dodig and Smith proved that the time complexity of the iGCT algorithm is O(n^4) [4].

This proves that the time complexity of the "apple carving" algorithm is O(n^2) + O(n^4) = O(n^4).

E. Space Complexity

Theorem 3 The space complexity of the "apple-carving" algorithm is O(n).

Proof: Number of points in a planar point set S is defined as n. The number of triangles in any full triangulation of S is known to be 2n – h – 2, where h is the number of points belonging to CH(S) [17]. The number of edges in any full triangulation of S is known to be 3n – h – 3, where h is the number of points belonging to CH(S) [17]. This implies that the variables in "apple carving" algorithm tracking both visited cities and candidate edges cannot have the space complexity greater than O(n).

This proves that the space complexity of the "apple carving" algorithm is O(n).

IV. Experimental Methodology

A. Objective

Our experimental objective was to test the validity of the proposed tour construction algorithm experimentally by analyzing how well the length of the resulting Hamiltonian cycle approximates the length of the optimal TSP.

B. Hypothesis

We hypothesize that the "apple carving" algorithm will outperform the traditional Convex Hull algorithm. We further hypothesize that the "apple carving" algorithm will outperform most of the traditional tour construction heuristics.

C. Data Sets

To perform our experiments, we selected 18 problem sets from TSPLIB, a well-known online problem library created to provide researchers with a broad set of test problems from various sources and properties for which the optimal TSP solutions are known [18]. We have chosen 11 problem sets which are given with points in general position (att48, *berlin52*, *ch130*, *eil51*, *eil76*, *eil101*, *gr96*, *gr137*, *rat99*, *rat195*, *rd100*). This was important as point sets in general position do not have 3 or more co-linear points. We have also chosen 7 problem sets with a significant number of co-linear points (*lin105*, *pr76*, *pr107*, *pr124*, *pr136*, *pr144*, *u159*). This was done to test performance of our framework in both point set configurations.

D. Programming

To achieve our experimental objectives we have programmed iGCT Algorithm in VBA for Excel. This algorithm takes a planar point set as an input, and produces a Hamiltonian cycle of S as an output. It also calculates the length of P'' found by "apple carving" algorithm in order to compare to the optimal TSP lengths for each of the problems in our problem set. All of our experiments were performed on Latitude 5490 laptop with Intel Core i5-8250U CPU @ 1.60GHz with 8GB of RAM, running Windows 10 64-bit operating system.

V. Results

Experimental results for 18 given problem sets can be found in Table I.

On average, polygons produced by the "apple carving" algorithm in our test problems are 8.1% longer than optimal TSP solutions. For *gr137* problem, the absolute error is the lowest at 1.9%, and for *pr124* problem, the error is the highest recorded at 15.9%. If we exclude point sets of 3 or more co-linear points, the absolute error drops to the average of 6.1%, with the maximum error recorded for *ch130* problem at 11.2%.

TABLE I. EXPERIMENTAL RESULTS

Set	S	TSP	P''	$\varepsilon^{A}_{P''}$	Co-linear
1	*att48*	108,159	118,702	9.8%	
2	*berlin52*	7,544	7,711	2.2%	
3	*ch130*	6,111	6,793	11.2%	
4	*eil51*	430	453	5.3%	
5	*eil76*	545	578	5.9%	
6	*eil101*	642	701	9.2%	
7	*gr96*	512	534	4.3%	
8	*gr137*	729	743	1.9%	
9	*lin105*	14,383	15,207	5.7%	Yes
10	*pr76*	108,159	112,152	3.7%	Yes
11	*pr107*	44,301	49,653	12.1%	Yes
12	*pr124*	59,030	68,069	15.3%	Yes
13	*pr136*	96,770	108,573	12.2%	Yes
14	*pr144*	58,535	67,867	15.9%	Yes
15	*rat99*	1,219	1,265	3.7%	
16	*rat195*	2,333	2,517	7.9%	
17	*rd100*	7,910	8,426	6.5%	
18	*u159*	42,075	47,354	12.6%	Yes

VI. CONCLUSIONS AND NEXT STEPS

We have introduced a simple algorithm that takes a full triangulation (iGCT) of a planar point set and reduces it to a simple polygon by removing triangles with low Triangle Inequality Measure starting from triangles on the convex hull of this point set. We have proved that the time complexity of the "apple carving" algorithm is O(n^2). We have also shown that the space complexity of the algorithm to be O(n). We have then demonstrated that, on average, polygons produced by this "apple carving" algorithm in our test problems are 8.1% longer than optimal TSP solutions. If we exclude point sets of three or more co-linear points, the absolute error drops to the average of 6.1%, with the maximum error recorded at 11.2%.

Based on these results and our literature review we conclude that "apple carving" algorithm produces better quality of solutions than any other construction heuristics other than match-twice-and-stitch heuristic, as evident in Fig. 5. Here it is important to note that the "apple carving" average results have been adjusted up by 0.8%, since HK lower bound is on average 0.8% lower than the optimal TSP solution [15].

Our initial research hypothesis that the "apple carving" algorithm will produce results superior to that of the classical Convex Hull Algorithm were met (9% average error for "apple carving" versus 12% average error for Convex Hull algorithm). We were also able to demonstrate that the "apple carving" algorithm performs significantly better than all the classical tour construction heuristics and is only slightly outperformed by Match-twice-and-stich heuristic introduced in 2004 [14].

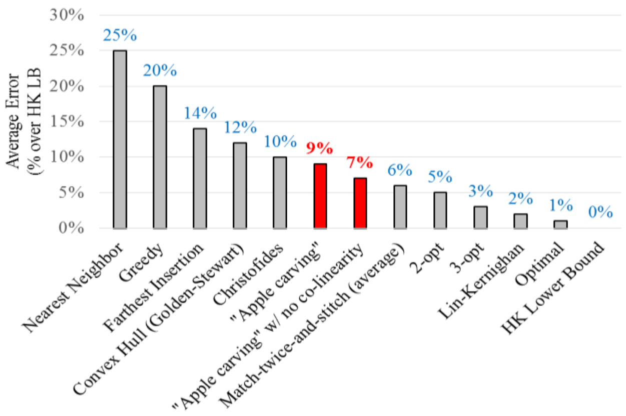


Fig 5. Typical Performance of Cited Heuristics over HK Lower bound (Including "Apple Carving" Algorithm Results).

Limitations in our work lie in the number of TSPLIB instances we used (i.e. 18 problems), as well as in the relatively small problem sizes employed (i.e. maximum of 195 points). To improve quality of our experiments we intend to expand our tests to all named TSPLIB instances, which will also allow us to compare how our algorithm performs on problems of varying size.

Finally, our future work will focus on fine-tuning the "apple carving" algorithm and adding the improvement steps of switching the relevant triangles in and out of the solution polygon depending on whether adding or removing related triangle pairs will result in desired tour improvements. Triangle pairs would be relevant and suitable for "swapping" in and out of the resulting polygon if the share at least one point, and their "swap" would not result in a loss of solution feasibility.

REFERENCES

[1] G. Reinelt, "Fast heuristics for large geometric traveling salesman problems" ORSA Journal on Computing, pp. 206-217, 1992.

[2] W. Stewart, Euclidean traveling salesman problems and Voronoi diagrams. School of Business Administration, College of William and Mary, 1997.

[3] A. N. Letchford, and N. A. Pearson, "Good triangulations yield good tours", Computers and Operations Research, vol. 35(2), 2008, pp. 638-647.

[4] M. Dodig, and M. Smith, "Novel heuristic for approximating minimum weight triangulation of planar point sets", unpublished.

[5] W. Cook, In Pursuit of the Traveling Salesman. Princeton University Press, 2012.

[6] R. Bellman, "Dynamic programming treatment of the travelling salesman problem", Journal of the ACM, vol. 9(1), pp. 61-63, 1962.

[7] R. Garey, D. Johnson, and R. Tarjan, "The Planar Hamiltonian Circuit Problem is NP-Complete", SIAM Journal on Computing, pp. 704-714, 1976.

[8] M. Held, and R. Karp, "A Dynamic Programming Approach to Sequencing Problems", Journal for the Society for Industrial and Applied Mathematics, vol. 10(1), pp. 196-210, 1962.

[9] G. Kizilates, and F. Nuriyeva, "On the nearest neighbor algorithms for the traveling salesman problem", Advances in computational science, engineering, and information technology, vol. 225, pp. 111-118, 2013.

[10] C. Nilsson, Heuristics for the Traveling Salesman Problem. Linköping University, 2003.

[11] D. Rosenkrantz, R. Stearns, and P. Lewis, "Approximate algorithms for the traveling salesperson problem", 15th Annual Symposium on Switching and Automata Theory, pp. 33-42, 1974.

[12] B. Golden, L. Bodin, T. Doyle, and W. Stewart Jr, "Approximate traveling salesman algorithms", Operations Research, vol. 28(3), pp. 694-711, 1980.

[13] N. Christofides, "Worst-case analysis of a new heuristic for the travelling salesman problem (No. RR-388)", Carnegie-Mellon University, Management Sciences Research Group, 1976.

[14] A. Kahng, and S. Reda, "Match twice and stitch: a new TSP tour construction heuristic", Operations Research Letters, pp. 499-509, 2004.

[15] D. Johnson, L. McGeoch, and E. Rothberg, "Asymptotic Experimental Analysis for the Held-Karp Traveling Salesman Bound", Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 341-350, 1996.

[16] R. Osserman, "The Isoperimetric Inequality", Bulletin of the American Mathematical Society, vol. 84(6), pp. 1182-1238, 1978.

[17] T. Vassilev, Optimal Area Triangulations. University of Saskatchewan, 2005.

[18] G. Reinelt, "TSPLIB - a traveling salesman problem library", INFORMS Journal on Computing, pp. 376–384, 1991.

Real-Time Cryptocurrency Price Prediction by Exploiting IoT Concept and Beyond: Cloud Computing, Data Parallelism and Deep Learning

Ajith Premarathne[1], Malka N. Halgamuge[2*], R. Samarakody[3], Ampalavanapillai Nirmalathas[4]
Velrada Capital Pty Ltd, Melbourne VIC 3000[1]
Department of Electrical and Electronic Engineering, The University of Melbourne, VIC 3010, Australia[2, 3, 4]

***Abstract*—Cryptocurrency has as of late pulled in extensive consideration in the fields of economics, cryptography, and computer science due to it is an encrypted digital currency, peer-to- peer virtual forex produced using codes, and it is much the same as another medium of the trade like real cash. This study mainly focuses to combine the Deep Learning with Data parallelism and Cloud Computing Machine learning engine as “hybrid architecture” to predict new Cryptocurrency prices by using historical Cryptocurrency data. The study has exploited 266,776 of Cryptocurrency prices values from the pilot experiment, and Deep Learning algorithm used for the price prediction. The four hybrid architecture models, namely, (i) standalone PC, (ii) Cloud computing without data parallelism (GPU-1), (iii) Cloud computing with data parallelism (GPU-4), and (iv) Cloud computing with data parallelism (GPU-8) introduced and utilized for the analysis. The performance of each model is evaluated using different performance evaluation parameters. Then, the efficiency of each model was compared using different batch sizes. An experimental result reveals that Cloud computing technology exposes new era by performing parallel computing in IoT to reduce computation time up to 90% of the Deep Learning algorithm-based Cryptocurrencies price prediction model and many other IoT applications such as character recognition, biomedical field, industrial automation, and natural disaster prediction.**

Keywords—Internet of things; IoT; data parallelism; deep learning; cloud computing

I. Introduction

Cryptocurrency is a technology dominant innovative form of digital currency that secures the financial transactions using cryptography, whereas concealing the identities of its users and minimize the counterfeit of the transactions. Cryptocurrency uses decentralized digital currency control that applies the distributed ledger technology, typically a Blockchain. The blockchain can be a distributed public financial transaction database, a public ledger or digital events that executed and shared between the participating parties. Participants in the Cryptocurrencies market build trust relationships through the formation of Blockchain supported cryptography techniques using hash functions. In 2008, an unknown group or an individual published a paper by introducing themselves under the name of Satoshi Nakamoto and paper entitled Bitcoin: A Peer-To-Peer Electronic Cash System”. This paper explains peer-to-peer online electronic cash payment system that would allow sending payments directly from one party to another without involving a financial institution. Bitcoin is the first realized Cryptocurrency concept created in 2009 and thought it extremely popular in 2017 [1]. The price of the Bitcoin has occasionally increased and therefore the value of the bitcoin is considered volatile. Hence, numerous economical entities try to predict the bitcoin price using different tools. This significant price movements of the bitcoin imply the requirement of accurate cryptocurrencies price prediction model to uphold the consistent economic policy. Thus, the demand for the cryptocurrencies price prediction mechanism is high. The cryptocurrencies price prediction model is prevalent around the world because most of the traders in the world use Cryptocurrencies to earn profits in an online market.

The blockchain databases have ready availability a large volume of data however the challenge is analyzing and storing this large volume of data on a time scale. Then, the cloud computing, which is the latest technological evolution of computational science, allowing groups to host, store process, and analyze large volumes of multidisciplinary data. Cloud computing is an internet-based utility service that provides virtualized service, storage, and databases, etc. The cloud technology is a distributed technology platform that leverage to provide highly scalable and resilient environments. Correspondingly, cloud computing architecture supports for the scalability, virtualization, and storage of large volume of structured and unstructured data based on the unlimited resources on demand [2]. Therefore, cloud computing is considered an appropriate platform for deep learning analytics. Google is one of the examples for the major Cloud computing providers. Thus, this study has used the Google Machine Learning (ML) Engine, as a Deep learning computing engine because, the Google ML Engine is easy to instruct for scaled data in deep learning algorithm [3]. Deep Learning referred to as achieved significant scalability and stability and generalization of training on big data. It can develop a model that converts inputs to outputs by extracting complex and non-linear hierarchical features of training data [4]. The programs of data-parallel entails with a series of operations and functioning to identify the large structured data. However, the parallelism can be either implicit or explicit, and can be regular or irregular [5].

This study aims to combine the Deep Learning with cloud computing and data parallelism based on the IoT concept for the development of cryptocurrencies price prediction model.

*Corresponding Author

II. Related Work

Enormous studies conducted to develop models for the cryptocurrency prices prediction however, there is a considerable gap in the research on predicting cryptocurrency involve with the machine learning algorithms. Many cryptocurrencies price prediction studies [6, 7, 8, 9] performed using standalone computers. However, Geourgoula et al. [10] discussed the Bitcoin price determinants and implemented a sentiment analysis technique that supports vector machines. The author explained that the network hash-rate and the frequency of the Wikipedia views had a significant positive correlation with the fluctuation of the Bitcoin price.

Greaves et al. [11] predicted the Bitcoin price by analyzing the Blockchain using SVM and ANN. The author reported that a regular ANN has 55 percent of price prediction accuracy. The study concluded that exchanges on the outside of the realm of the Blockchain have technically dictated price and it limited the Blockchain data predictability. Similarly, Matta et al. [12] studied the effect of tweets on Twitter and Trend views of Google for the price of Bitcoin with 60 days as sample size and sentiment as a variable. The author found that both Google Trend views and positive tweets have moderately correlated to the Bitcoin price fluctuation and that correlation can be used to predict the cryptocurrencies price. However, the inadequate sample size is a major drawback of the study and prediction based on the social media comments may not be a reliable source for the scientific studies. Steinkrau et al. [13] implemented a GPU-based ANN model and reported that the model is three times faster training and testing than a CPU. Ciresan et al. [14] also reported that GPU-based deep natural network training is forty times faster than a CPU for the image recognition. David Sheehan has proposed a Cryptocurrencies price prediction algorithm [6] based on Long Short-Term Memory (LSTM) neural network model. Correspondingly, Alex (2014) [15] suggested a method for paralleling the training of convolutional neural networks across multiple GPUs.

However, any of these studies did not exploit the IoT basic concepts and cloud computing phenomena in IoT along with Deep Learning for the Cryptocurrencies real-time price prediction. Therefore, this paper explains to quantify the impact of computation time of Deep Learning algorithm training on four models ((i) Standalone PC - (ii) GPU1 – without data parallelism model (iii) GPU4 – with data parallelism model and (iv) GPU8 – with data parallelism model) with a high accuracy percentage of Cryptocurrency price prediction. This study mainly focused on Parallel Processing and Cloud computing along with the internet of things (IoT) concept to develop a cryptocurrencies price prediction model.

Main contribution of the paper

- Real-time Cryptocurrency price was predicted by exploiting the Internet of things (IoT) concept and beyond.
- Data parallelism and Cloud Computing Machine learning engine were combined with Deep Learning and this hybrid architecture is applied to Cryptocurrency historical data to predict new Cryptocurrency price.
- Three hybrid architecture was developed for cryptocurrency data training and predicting purpose (i) standalone PC, (ii) Cloud computing without data parallelism (GUP-1), (iii) Cloud computing with data parallelism (GUP-4).
- Cloud computing technology secure new trends in performing parallel computing in IoT to reduce computation time up to 90% of the Cryptocurrency price prediction model using Deep Learning algorithm.
- Proposed hybrid architecture can be used in any application including in IoT applications such as character recognition, biomedical field, industry automation and natural disaster prediction.

The rest of the paper is organized as follows: Section II describes related work and the main contribution of the paper. Section III introduces how the data is collected and pre-processed and techniques to combine Deep Learning with cloud computing and data parallelism based on the IoT concept. Section IV provides results, and Section V provides related discussion. Finally, the paper concludes in Section VI.

III. Materials and Methodology

A. Data Collection and Preparation

Historical Cryptocurrencies data from the Quandal database collected and recorded daily for four years at different time instances. Then, the data normalized by implementing Min-Max Scalar technique and smoothened over the complete period and normalized data were retrieved up to a current date subsequently. Data preparation performed before the training process by using deep learning algorithm.

Before training the network, the data set scaled to converge the system efficiently. Then, the scaled data set divided into two sets as “training data set” and “testing data set”. The deep learning algorithm trained using the training data set and accuracy of the Cryptocurrencies price prediction for an unseen data tested using the testing data set.

The testing data set that manipulated to predict the Cryptocurrencies price trained by creating Neural Network Model which has Five-layers including input, output, and three hidden layers. The ReLU activation function applied for the hidden layers as it can increase the training efficiency. The Liner activation utilized for an output layer as it can pass values without any modification. Then, update the quality and speed of the model parameters using SGD optimizer.

B. Training Methods

Mean Squared Error (MSE), Mean Absolute Error (MAE), variance and computation time (CPU processing time) computed for each model to identify the best-fitted model to prediction of Cryptocurrencies price.

C. Performance Evaluation of the Four Models by Comparing the Batch Size

The MAE values, MSE value, Explained variance Score, Accuracy of the prediction (R2), Min-Max Scalar, Efficiency

Comparison Percentage and Efficiency of Computation Time of each hybrid architecture model compared using five different size of datasets (batch size) such as (i) Batch 32 (ii) Batch 256*4 (iii) Batch 256*8 (iv) Batch 256*16 and (v) Batch 256*32. The experiment performed for 50, 100, 200, 500, 1000 and 5000 epochs.

D. *Cryptocurrency Price Prediction Mechanism*

Fig. 3 describes the Cloud computing connected Cryptocurrencies predicting mechanism process.

Following steps (Fig. 1) explained the detail procedure for the Cloud computing connected Cryptocurrencies predicting mechanism.

Step 1: Retrieve historical Cryptocurrency data from the internet and save as CSV file.

Step 2: Load historical Cryptocurrency data to the desktop computer.

Step 3: Scale historical Cryptocurrency data to between 0-1 and then save back as CSV file in the desktop computer.

Step 4-1: Train the Cryptocurrency prediction model using deep learning algorithm without the Cloud computing model.

Step 4-2: Train the Cryptocurrency prediction model using deep learning algorithm in with Cloud computing model or with parallel Cloud computing.

Step 5-1: Save the trained Cryptocurrency prediction model in Cloud computing.

Step 5-2: Save the trained Cryptocurrency prediction model in Cloud computing.

Step 5: Retrieve real-time Cryptocurrency data from the internet as CSV file data.

Step 6: Feed lives Cryptocurrency data to train the Cryptocurrency prediction model that saved in Cloud computing.

Step 7: Get the result back from Cloud computing and show the predicted Cryptocurrency price.

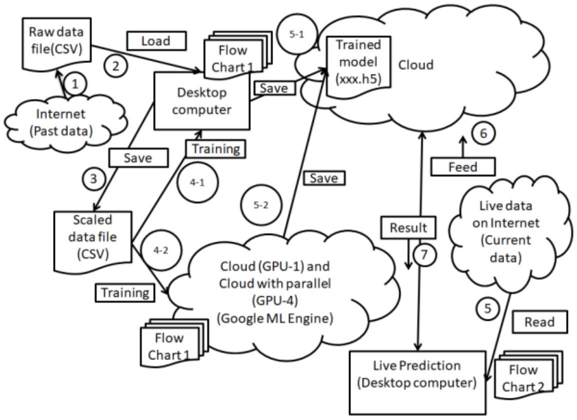


Fig. 1. Cloud Computing Connected Cryptocurrencies Predicting Mechanism using the Deep Learning Algorithm.

E. *Deep Learning Training Model in with or without Cloud Computing*

Flow Chart 1 emphasizes in Fig. 5 describes the training phase of the Cryptocurrencies price prediction using deep learning algorithm.

F. *Client-Side Cryptocurrency Price Predicting Model*

The Flow Chart 2 shown in Fig. 3 and Fig. 4 describes the prediction algorithm which used for the training method. Then, the training method saved on the Cloud. Finally, this training method used to predict the Cryptocurrencies price for unseen newly arrived data.

G. *Data Parallelism Cloud Computing Working Methodology*

The data parallel method explained by [16] has practiced for parallel training as showed in Fig. 2 and steps are as followed.

Step 1: Dataset was divided into eight datasets

Step 2: Feed those data sets into four graphics processing units (GPUs)

Step 3: Each GPU computes different data set of the batches.

Step 4: Data parallelism used synchronization between model parameters and model parallelism doing synchronizing between input and output values between the data chunks.

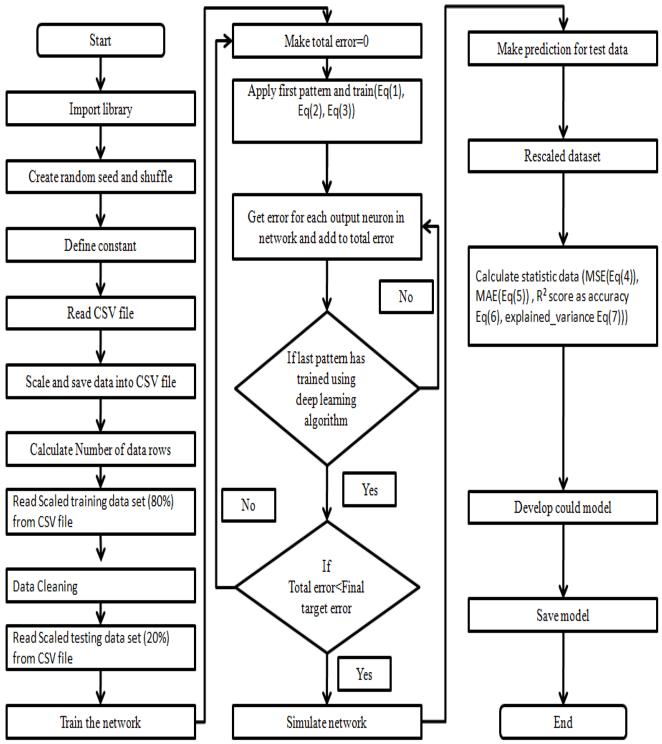

Start
Import library
Create random seed and shuffle
Define constant
Read CSV file
Scale and save data into CSV file
Calculate Number of data rows
Read Scaled training data set (80%) from CSV file
Data Cleaning
Read Scaled testing data set (20%) from CSV file
Train the network
Make total error=0
Apply first pattern and train(Eq(1), Eq(2), Eq(3))
Get error for each output neuron in network and add to total error
No
If last pattern has trained using deep learning algorithm
Yes
No
If Total error<Final target error
Yes
Simulate network
Make prediction for test data
Rescaled dataset
Calculate statistic data (MSE(Eq(4)), MAE(Eq(5)) , R² score as accuracy Eq(6), explained_variance Eq(7)))
Develop could model
Save model
End

Fig. 2. Server Side: The Cryptocurrencies Price Prediction Training Flow Chart using Deep Learning Algorithm (Flow Chart 1).

H. Algorithms

Two algorithms developed for the prediction of Cryptocurrencies price. The algorithm 1 used to compute the MSE, MAE, R2 and explained variance of the historical data and to develop a Cloud computing training model. Then, the training model is developed by the Algorithm 1 (that has saved on the Cloud computing) is used for the Algorithm 2 to predict the Cryptocurrencies price of the live data.

Algorithm 1: Cryptocurrencies price prediction training using Deep Learning algorithm

```
Begin
Import library
Create random seed and shuffle
Define constant
Read CSV file
Scale and save data into CSV file
Calculate Number of data rows
Read Scaled training data set (80%) from CSV file
Data Cleaning
Read Scaled testing data set (20%) from CSV file
Train the network
While total errors ==0:
        Apply the first pattern and train the network
        Get error for each output node in the network and add to the total error
If the last pattern has trained, then:
If total error < final target error, then:
        End training
                End If
        End If
End While
Simulate network
Make a prediction for test data
Rescaled dataset
Calculate statistic data (MSE), MAE, $R^2$ score as Cryptocurrency price prediction accuracy, explained variance)
Develop Cloud computing model
Save Cloud computing model
End
```

Algorithm 2: Cryptocurrencies price predicting using a Deep Learning algorithm

```
Begin
Import library
Define constant
Initialize the variable
Initialize the plot
Initialize the Google Credentials Variable
While True:
        Read live data from the server
        Scaled data
        For $j$ in range (0, length of the data file):
                Assign Cloud computing input data
                Read credential file
                Gets prediction from Cloud computing
                Save on data frame
        End for
        Plot the live prediction graph
        Wait for new data
End while
End
```

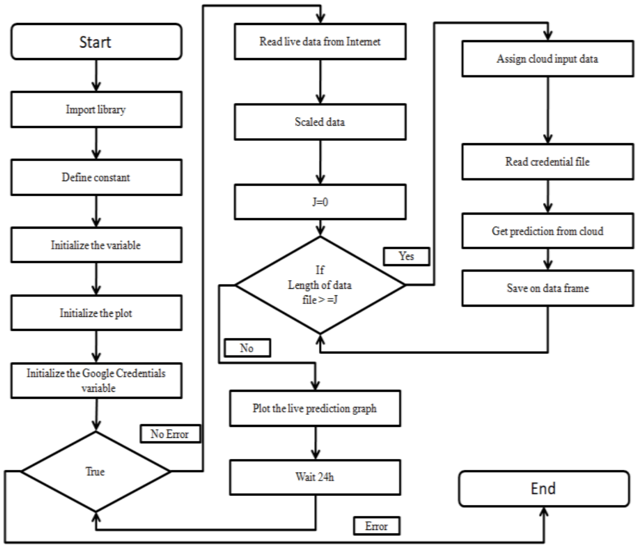


Fig. 3. Client Side: Cryptocurrencies Predicting Flow Chart using Deep.

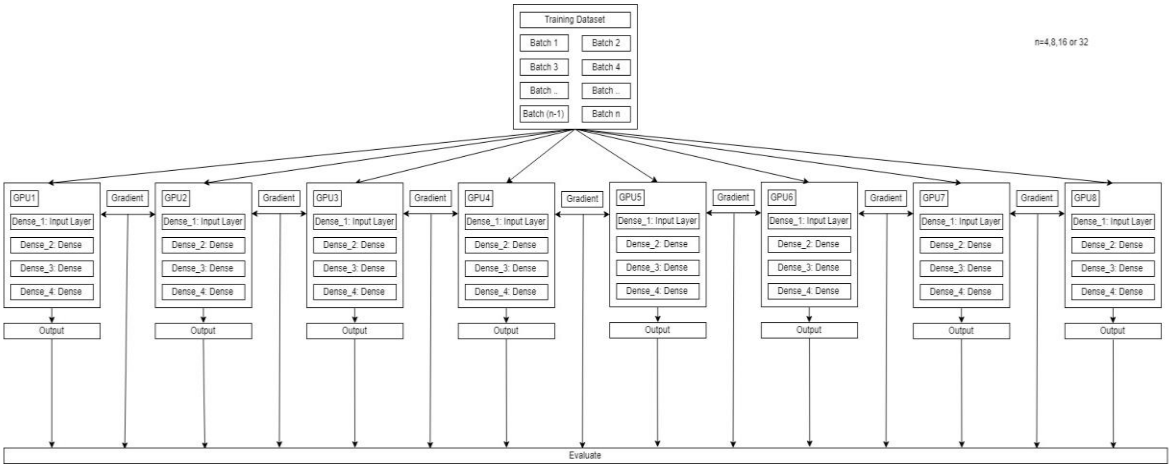


Fig. 4. The Cryptocurrency Data Parallelism Training Block Diagram using Deep Learning Algorithm in Cloud Computing Learning Algorithm (Flow Chart 2).

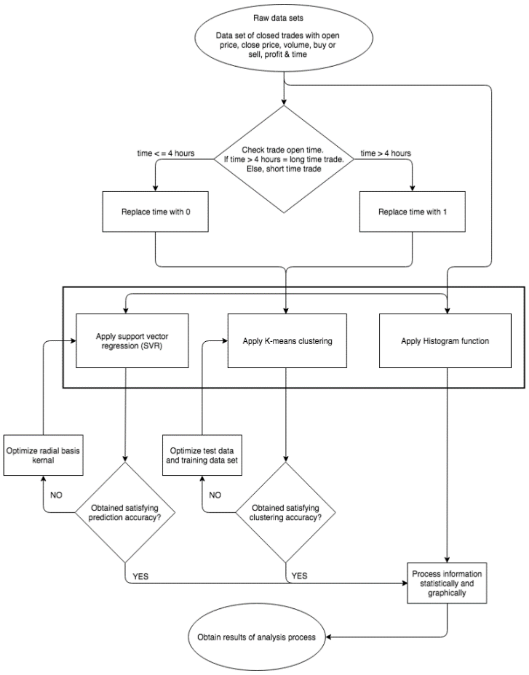

Raw data sets
Data set of closed trades with open price, close price, volume, buy or sell, profit & time
Check trade open time. If time > 4 hours = long time trade. Else, short time trade
time < = 4 hours
time > 4 hours
Replace time with 0
Replace time with 1
Apply support vector regression (SVR)
Apply K-means clustering
Apply Histogram function
Optimize radial basis kernal
Optimize test data and training data set
NO
NO
Obtained satisfying prediction accuracy?
Obtained satisfying clustering accuracy?
YES
YES
Process information statistically and graphically
Obtain results of analysis process

Fig. 5. Overall Methodology (Flow Chart 3).

I. Comparison of the Efficiency Results

The efficiency of the model compared using Efficiency Comparison Percentage (ECP) equation and Table I describes the parameters and equation for each step.

Finally, overall methodology has drafted as showed in Fig. 7.

TABLE. I. HYBRID TECHNIQUE: CRYPTOCURRENCY HISTORICAL DATA TRAINING METHODS

Method	**CPU**			**Memory (GB)**
Standalone PC	Intel core i3 – 7100U - 2.4 Hz			8
Cloud computing Method	GPU name	GPU model	GPUs	GPU memory (GB) (GDDR5)
Cloud computing without data parallelism (GPU1 – without data parallelism model)	*Optimizer SGD (Stochastic gradient descent):* In each training, SGD will update the parameter Standard_GPU	NVIDIA Tesla K80	1	12
Cloud computing with data parallelism (GPU4) (GPU4 – with data parallelism model)	Complex_model_l _GPU	NVIDIA Tesla K80	4	48
Cloud computing with data parallelism (GPU8) (GPU8 – with data parallelism model)	Complex_model_l _GPU	NVIDIA Tesla K80	8	120

IV. RESULT

A. Performance Evaluation of the Four Models by Comparing the Batch Size

1) Comparing the performance evaluation of the standalone PC method: According to Fig. 6 the batch 32 recorded 88.706 of the highest prediction accuracy value in epoch 5000 and it consumes 765.690 minutes while the batch 256*32 recorded 35.886 as the lowest accuracy rate in epoch 50 during 41.552minutes. According to the results of these comparisons, the highest prediction accuracy value observed from Batch 256*4 as 85.646 while it consumed 81.030 minutes in epoch 5000. However, Batch 256*16 has significant prediction accuracy of 81.266 and efficiency is 59.78 minutes.

2) Comparing the performance evaluation of the GPU1 – without data parallelism: Fig. 7 indicates that the maximum and minimum prediction accuracy values of the GPU1 – without data parallelism models observed in epoch 5000 and 50 respectively for all batch sizes. However, batch 32 recorded, 88.703 as maximum prediction accuracy value and consume 1785.97 minutes. The batch 256*32 had 35.874 as minimum prediction accuracy value and it used 7.624 minutes. Conferring to the results in Fig. 7, epoch 5000 reported 85.647 of prediction accuracy as the highest value in Batch 256*4 while it consumes 152.983 minutes to fulfill the target efficiency. However, for the GPU1 – without data parallelism model the Batch 256*16 reached 81.267 accuracy percentage.

3) Comparing the performance evaluation of the GPU4 – with data parallelism: Fig. 8 emphasizes the accuracy value comparison of the five batches. According to the result, the Batch 32 has 87.779 of the highest prediction accuracies in epoch 5000 while the Batch 256*32 has 35.874 of prediction accuracy which is reported as the lowest.

4) Comparing the performance evaluation of the GPU8 – with data parallelism: The highest prediction accuracy of 87.071 reported by the batch 32 in epoch 5000 and consumed 686.541 minutes (Fig. 9). However, the batch 256*32 has the best efficiency which is 28.23 minutes and prediction accuracy of 79.088 for the GPU8 – with data parallelism model.

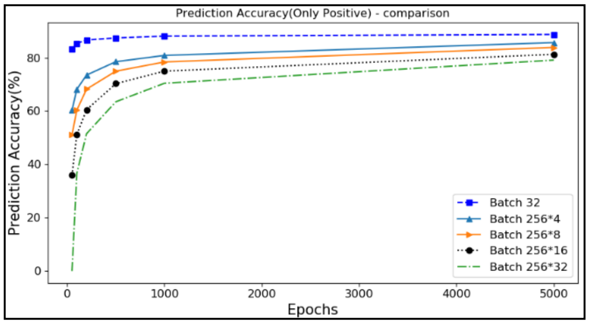


Fig. 6. Comparing the Prediction Accuracy Values of the Standalone PC Method Related to different batch Size.

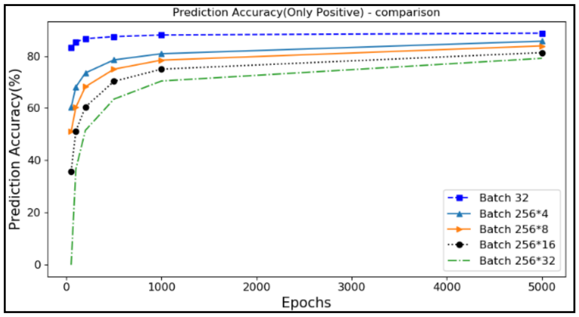


Fig. 7. Comparison of the Prediction Accuracy of the GPU1 – without Data Parallelism whereas; Parallelism – Efficiency (without Batch 32) Values Related to different Batch Size.

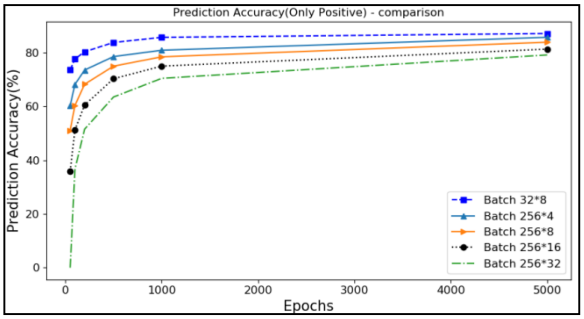


Fig. 8. Comparison of the – Prediction Accuracy Values of the GPU4 – with Data Parallelism Related to different Batch Size.

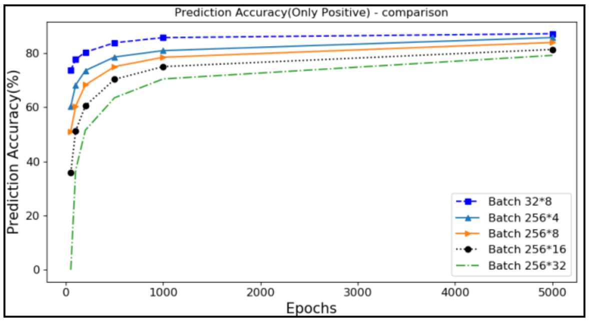


Fig. 9. Comparison of the Prediction Accuracy Values of the GPU8 – with Data Parallelism Related to different Batch Size.

B. *Comparison of the Efficiency Percentage Results*

Efficiency percentage of each model was compared using the methodology described in the Table II.

Method A; Cloud computing without data parallelism (GPU1) model vs. Cloud computing with data parallelism model (GPU4).

Method B; Cloud computing without data parallelism (GPU1) model vs. Standalone PC model.

Method C; Cloud computing without data parallelism (GPU1) model vs. Cloud computing with data parallelism model (GPU8).

Method D; Standalone PC model vs. Cloud computing with data parallelism model (GPU4).

Method E; Standalone PC model vs. Cloud computing with data parallelism model (GPU8).

Method F; Cloud computing with data parallelism (GPU4) model vs. Cloud computing with data parallelism model (GPU8).

*1) Efficiency comparison of the algorithms for batch 256*8:* Fig. 10 reveals that the EPC results from Method A to Method F for the Batch 256*8. The GPU1 has no data parallelism, therefore, it spent a lot of time on the training compared to the GPU4. However, the ratio of Method A is significantly higher than Method B, Method C, Method D, and Method F. In Method B, the GPU1 model runs in cloud platform and the Standalone PC without cloud just like a laptop computer. The GPU1 module consumed more time for the training compared with Standalone PC model because the GPU1 module requires considerable time to flush the memory. The Standalone PC model has higher efficiency percentage from 50 to 200 epochs while runs faster within that epochs range than the GPU4 model. Subsequently, the efficiency percentage of the Standalone PC model slightly slower than GPU4 model. Hence, until 200 epochs Method B ratio is higher than Method A. The GPU1 model in Method C took more time for the training the GPU8 model because it does not include the data parallelism. However, GPU8 model is slightly slower than the GPU4 and Standalone PC models; thus, the ratio of the Method C comparatively lower than Method A, Method B, and Method F.

TABLE. II. COMPARISON EQUATIONS FOR EFFICIENCY COMPARISON PERCENTAGE (ECP)

Comparison Method	Description	The equation for Efficiency Comparison Percentage (ECP) calculation
Method A	Cloud computing without data parallelism (GPU1) model vs. Cloud computing with data parallelism model (GPU4)	Percentage = [(Cloud computing without data parallelism (GPU1)- Cloud computing with data parallelism (GPU4)) /Cloud computing (GPU1)] * 100 %
Method B	Cloud computing without data parallelism (GPU1) model vs. Standalone PC model	Percentage = [(Cloud computing (GPU1) - standalone PC) / Cloud computing (GPU 1)] * 100 %
Method C	Cloud computing without data parallelism (GPU1) model vs. Cloud computing with data parallelism model (GPU8)	Percentage = [(Cloud computing without data parallelism (GPU1)- Cloud computing with data parallelism (GPU8)) /Cloud computing (GPU1)] * 100 %
Method D	Standalone PC model vs. Cloud computing with data parallelism model (GPU4)	Percentage = [(Standalone PC - Cloud computing with data parallelism (GPU4)) / Standalone PC] * 100 %
Method E	Standalone PC model vs. Cloud computing with data parallelism model (GPU8)	Percentage = [(Standalone PC - Cloud computing with data parallelism (GPU8)) / Standalone PC] * 100 %
Method F	Cloud computing with data parallelism (GPU4) model vs. Cloud computing with data parallelism model (GPU8)	Percentage = [(Cloud computing with data parallelism (GPU8)- Cloud computing with data parallelism (GPU4)) /Cloud computing (GPU8)] * 100 %

Efficiency Percentage comparison results of Method D indicated that the Standalone PC model required more time for the training compared to GPU4 model. In Method D, the efficiency of both GPU4 and Standalone PC models faster than the GPU1 and GPU8 models. Therefore, Method D ratio is the lowest due to comparing the two fastest algorithms. According to the equation of Method E, Standalone PC model is faster than GPU8 model; hence, the Method E ratio is negative and Method E line is not plot in Fig. 10. The GPU8 model in Method consumed considerable time for the training compared to the GPU4 model. The inter-process communication of the GPU8 model may be the reason for this substantial time consumption and as a result Method F ratio is well above Method C and Method D.

*2) Efficiency comparison of the algorithms in batch 256*16:* The GPU1 model in Method A has no data parallelism therefore it spent lot time for the training than the GPU4 model and the efficiency percentage of the GPU1 model in Batch 256*16 is significantly slower than the Batch 256*8 (Fig. 11). However, the ratio of Method A is higher than the other methods. As in Batch 256*8 for Method B, the GPU1 model spent more time on the training than the Standalone PC model because the GPU1 model consumed considerable time to flush the memory. Results of the Method B in Batch 256*16 has evidently shown that the Standard PC model is slower than the GPU4 and GPU8 model and therefore, the ration of the Method B tracked below the Method A and Method C. In Method C, the GPU8 model is slightly speed than the Standalone PC model, however, the efficiency percentage of the GPU1 model in Method C for the Batch 256*16 showed comparatively higher efficiency percentage than the Batch 256*8. Thus, the Method C ratio is slightly below than the Method.

Method D result illustrated that the Standalone PC model is 50% slower than the GPU4 model. Therefore, Method D ratio is lower than Method A, Method C, and Method B while higher than the Method E and Method F. The efficiency ratio of the Standalone PC model in Method E is slightly slower than the GPU8 model hence, Method E ratio is above Method F. The GPU8 model in Method F consumed substantial time for the training than the GPU4 model. It caused to slower the GPU8 model and inter-process communication may be the reason for this significant time consumption. As a result, the Method F ratio is lowest for the Batch 256*16.

*3) Efficiency percentage comparison of the algorithms in batch 256*32:* In Method A, GPU1 spent significant time on the training compared to the GPU4 model (Fig. 12). However, the ratio of Method A is significantly higher than other methods. The GPU1 model in Method B runs in cloud platform and it consumed more time for the training compared with the Standalone PC model because the GPU1 model take some time to flush the memory. The efficiency ratio of the GPU8 model in Method C is slightly speeding than the Standalone PC model, therefore, the ratio of the Method C is lower than the Method A and higher than the Method B, Method D, Method E, and Method F.

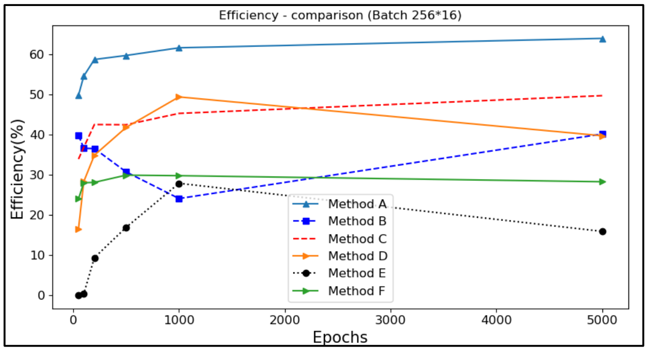


Fig. 10. Efficiency Comparison Percentage for the Batch 256*8.

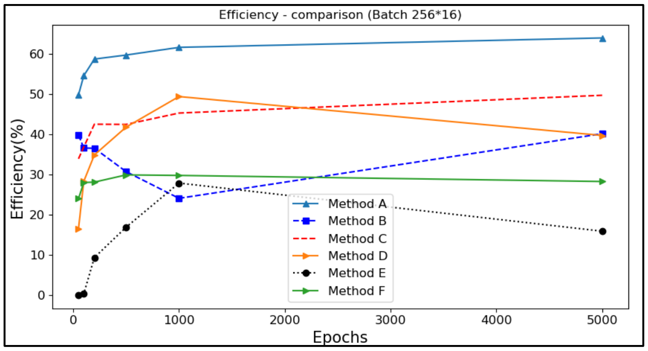


Fig. 11. Efficiency Percentage Comparison for the Batch 256*16.

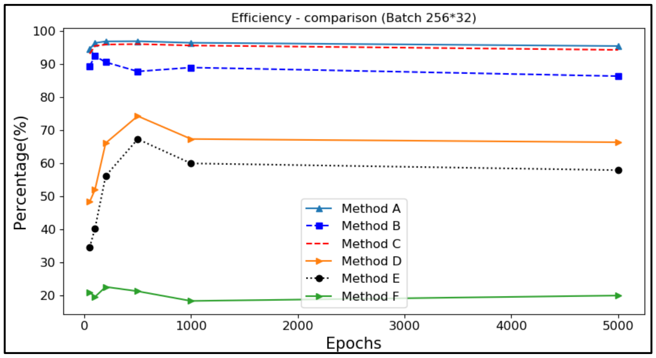


Fig. 12. Efficiency Percentage Comparison for the Batch 256*32.

Both GPU4 and Standalone PC models in Method D are faster than the GPU1 and GPU8 models. When considering Method D in Batch 256*32 and Batch 256*8, the efficiency percentage of the GPU4 model in Batch 256*32 is higher than the Batch 256*8. When consider Method E for Batch 256*32 the Standalone PC model is noticeably faster than the GPU8 model hence, Method E ratio is the lowest ratio for the Batch 256*32. The GPU8 model in Method F required more time for the training compared to the GPU4 model while results of the Method F in Batch 256*32 is faster than Batch 256*8. However, the ratio of Method F is higher than Method E for the Batch 256*32.

V. Discussion

This study aims to predict real-time Cryptocurrency Price by using Deep Learning algorithm whereas exploiting IoT concepts and beyond using Cloud computing and Data Parallelism. Time consumption is the major barrier for the training a large data set sequenced in the neural network. Therefore, this application primarily concerned to develop an algorithm to forecast the Cryptocurrencies price prediction accuracy. Numerous research experts discussed cloud computing [15, 17], Deep learning algorithms [18], cryptocurrency price prediction, Bitcoin [6, 7, 9] and data parallelism [19, 20] separately as three different topics. Thus, it has a potential and significant correlation between these three approaches and can be experimented together to explain precise model for the Cryptocurrencies price prediction. However, this potential was ignorance and created a substantial gap in the field. Therefore, the experimental methodology of this study combined these three studies into a single platform to exploit the IoT basic concept for real-time Cryptocurrencies price prediction based on historical data. The main challenge of the real-time Cryptocurrencies price predicts models is that the application accuracy in real-world due to the fluctuating nature of the Cryptocurrencies. Similarly, identifying daily trends in the Bitcoin market while gaining insight into optimal features surrounding Bitcoin price is important because they try to predict the sign of the regular price change with the highest possible accuracy [9]. The Bayesian Neural Networks are a precise approach to estimate the maximum likelihood of Cryptocurrencies price and explaining the high volatility of the recent Bitcoin price [7]. Alternatively, reduce the training time of the Deep Learning algorithm is a noteworthy challenge for the cryptocurrency price prediction approaches. However, without Cryptocurrencies price prediction accuracy, computation time useless.

This study identified three major gaps in the cryptocurrencies price prediction models through the literature review as (1) accuracy of the application (2) long computation time and (3) application of IoT concepts to the prediction models. Concerning all the gaps in the current Cryptocurrencies price prediction applications, this study developed four hybrid architecture models, namely, (i) Standalone PC, (ii) GPU1 – without data parallelism model, (iii) GPU4 – with data parallelism model, and (iv) GPU8 – with data parallelism model for Cryptocurrencies training methods with a similar deep learning algorithm.

Primarily, the study concerned to enhance the accuracy of the Cryptocurrencies price prediction model and suggests an alternative to overcome the factors effect to reduce the prediction accuracy using Deep learning algorithm. The study utilized 266,776 historical data for the training of Cryptocurrencies price prediction Deep learning algorithm. The experiment has maintained maximum epoch for the Deep learning algorithm training as 5000 because the study expected to achieve more than 80% of price prediction accuracy. This study applied IoT technology combined with the Cloud computing to predict Cryptocurrencies price and to train the Cryptocurrencies price prediction, model. Also, the volume of the data set considerably influence to the accuracy and computation time of the prediction models and thus used five different batch sizes for the experiment. The accuracy percentage of the prediction and volume of the data set has a positive correlation which means the prediction of the big data set can be higher compared to the small volume of data set. The Google ML engine provides different types of GPU for Cloud computing with data parallelism models which can be utilized for Deep Learning training. Therefore, types of GPU may have potential to reduce the computation time and accuracy of the training methods. The main advantage of the parallelism data is that it can be divided into a few batches to reduce the data set size of one batch, and then GPU can compute an individual quantity of the data set. However, the reduction of the volume of the data set that simulated to GPU affected to the prediction accuracy. Furthermore, the study identified that accuracy of the Cryptocurrency price prediction models can be increased using fully connected dense neural network with ReLU activation function in hidden layers and linear activation function in the output layer.

Deep Learning algorithm training is a highly time-consuming process when the data set is large. Subsequently, this study combined the IoT concept with parallel processing and Deep Learning to reduce the computation time of the prediction of the models by training the historical data over pre-determined time slots. Firstly, the Standalone PC model was trained, and highest prediction accuracy which 88.7% was obtained by Batch 32 within 765.69 min. The best accuracy percentage for this model was 81.27% and this could be achieved within 59.78 min. by the Batch 256*16. Secondly, the GPU1 – without data parallelism model was trained. The Batch 32 reported the highest prediction accuracy as 88.7% but it consumed 1785.97 min which is not practical. However, Batch 256*16 has the best efficiency which is 99.84 min and accuracy percentage was 81.27 for this model. Thirdly, the GPU4 – with data parallelism model trained and Batch 256*4 represented 87.78% of accuracy within 909.85min. For this model Batch, 256*32 has the best efficiency which is 22.58 min and accuracy was 79.09%. Finally, the GPU8 – with data parallelism model trained and 87.07% the highest accuracy percentage could be observed from Batch 32 within 686.54min. Batch 256*32 has the best efficiency which is 28.23 min for 79.09% of accuracy. All four models achieved almost 80% Cryptocurrencies price prediction accuracy.

The experimental results confirmed that the GPU4 – with data parallelism and the GPU8 – with data parallelism models can reduce the computation time which is approximately within 30 minutes for the large batch sizes. Few authors applied the Deep Learning approach with the parallel neural network [17], data parallelism [15] and Parallel Consensual Neural Networks [20] to reduce the computation time. Similarly, [19] has discussed the effect of traffic flow in cloud computing for the computation time using different types of parallel architectures. All these studies proved that a combination of Cloud computing with data parallelism for the training model significantly reduce the computation time. However, the data parallelism models can be executed for a large set of historical data, and Deep Learning training with the different GPU types available on the Google ML engine. Furthermore, proposed hybrid architecture models can be utilized in any IoT application. Correspondingly, future experiments can be focused on device parallelism with cloud computing (GPU-8) for the Deep Learning training. Besides, understanding decentralized approaches for big data databases [21, 22], decision making utilizing predicting techniques [23, 24], could be an inspiring method to make the Internet of Things into one of the future Fourth Industrial Revolution Technologies (4IR/FIR).

VI. Conclusion

This study trained four hybrid architecture models to predict real-time Cryptocurrencies price using deep learning algorithm by exploiting the IoT concept. The experimental results confirmed that Cloud computing technology stimulus to secure new trends by performing parallel computing in IoT. Similarly, the results of this study confirmed that data parallelism and Deep Learning algorithm-based Cryptocurrencies price prediction models can reduce computation time up to 90% with 80% of accuracy. However, the comparison between the model which did not train with data parallelism namely the Standalone PC and the GPU1 – without data parallelism models revealed the insignificant outcome. The Batch 256*32 in GPU8 – without data has the best accuracy which is 79.09%. The GPU4 – without data parallelism model resulted in similar results and the Batch 256*32 reported 79.09% of accuracy. These values revealed that the potential of Cloud computing with data parallelism (GPU-8 and GPU-4) models to use for Cryptocurrencies price prediction. Therefore, the experimental results concluded that uses of Cloud computing with data parallelism (GPU-4 and GPU-8) models can accelerate the Cryptocurrencies price prediction process than all other hybrid architecture models tested in this study and this may vary with the size of the batch. Ultimately, there is an enormous potential to apply the proposed hybrid architecture models into any other deep learning models such as character recognition, the biomedical field, in addition to any application in IoT such as industrial automation and natural disaster prediction.

Author's Contribution

A.P. and M.N.H. conceived the study idea and developed the analysis plan. A.P. analyzed the data and wrote the initial paper. M.N.H. helped to prepare the figures and tables and finalizing the manuscript. R.S. completed the final editing of the manuscript. All authors read the manuscript.

References

[1] A. Rosic, (2016). What is Cryptocurrency: Everything You Need to Know [Ultimate Guide]. Retrieved from Blockgeeks.com: https://blockgeeks.com/guides/what-is-cryptocurrency.

[2] H. Yan, P. Yu, D. Long (2019). Study on Deep Unsupervised Learning Optimization Algorithm Based on Cloud Computing. In 2019 International Conference on Intelligent Transportation, Big Data & Smart City, pp 679-681.

[3] Cloud ML Engine Overview. (2018). Retrieved from Google.com: https://Cloud computing.google.com/ml-engine/docs/tensorflow/ technical-overview

[4] L. Song, J. Mao, Y. Zhuo, X. Qian, H. Li, Y. Chen (2019). HyPar: Towards Hybrid Parallelism for Deep Learning Accelerator Array. In 2019 IEEE International Symposium on High Performance Computer Architecture, pp. 56-68.

[5] [G. Onoufriou, R. Bickerton, S. Pearson, G. Leontidis (2019). Nemesyst: A Hybrid Parallelism Deep Learning-Based Framework Applied for Internet of Things Enabled Food Retailing Refrigeration Systems. arXiv preprint arXiv:1906.01600.

[6] D. Sheehan, (2017). Predicting Cryptocurrency Prices with Deep Learning. Retrieved from Github: https://dashee87.github.io/deep%20learning/python/predicting-cryptocurrency-prices-with-deep-learning/

[7] H. Jang, J. Lee, "An Empirical Study on Modelling and Prediction of Bitcoin Prices with Bayesian Neural Networks based on Blockchain Information," IEEE Early Access Articles, vol. 99, pp. 1-1, 2017.

[8] S. McNally, "Predicting the price of Bitcoin using machine learning," School Comput., Nat. College Ireland, Dublin, Ph.D. dissertation 2016.

[9] S. Velankar, S. Valecha, S. Maji. (2018). Bitcoin Price Prediction using Machine Learning. International Conference on Advanced Communications Technology, 144-147, 11-14 Feb. 2018.

[10] I. Georgoula, D. Pournarakis, C. Bilanakos, D. Sotiropoulos, G. M. Giaglis (2015). Using time-series and sentiment analysis to detect the determinants of bitcoin prices, SSRN Electronic Journal.

[11] C. G. AkcoraAsim, K. Dey, A. Dey, Y. R. Gel, M. Kantarcioglu (2018) Forecasting Bitcoin Price with Graph Chainlets, Advances in Knowledge Discovery and Data Mining.

[12] Matta, M., Lunesu, I., & Marchesi, M. (2015). Bitcoin Spread Prediction Using Social and Web Search Media. In UMAP Workshops.

[13] D. Steinkrau, P. Y. Simard, and I. Buck (2005) Using GPUs for machine learning algorithms in Proceedings of the Eighth International Conference on Document Analysis and Recognition. IEEE Computer Society, pp. 1115-1119.

[14] Cireşan, D. C., Meier, U., Gambardella, L. M., & Schmidhuber, J. (2010). Deep, big, simple neural nets for handwritten digit recognition. Neural computation, 22(12), 3207-3220.

[15] A. Krizhevsky, (2014). One weird trick for parallelizing convolutional neural networks. arXiv:1404.5997, 1-7.

[16] M. Whitney, (2016). Deep Learning with Multiple GPUs on Rescale: Torch. Retrieved from Blog.rescale.com: https://blog.rescale.com/deep-learning-with-multiple-gpus-on-rescale-torch/

[17] F. Åström, R. Koker. (2011). A parallel neural network approach to prediction of Parkinson's Disease. Expert Systems with Applications, 12470-12474.

[18] A. A. Diro and N. Chilamkurti, "Distributed attack detection scheme using deep learning approach for Internet of Things", Future Generation Computer Systems, Volume 82, May 2018, Pages 761-768.

[19] P. Sekwatlakwatla, M. Mphahlele, T. Zuva. (2016). Traffic Flow Prediction in Cloud Computing. International Conference on Advances in Computing and Communication Engineering, 123-128, 28-29 Nov. 2016.

[20] J. Ekanayake, X. Qiu, T. Gunarathne, Scott Beason, Geoffrey Fox. (n.d.). High Performance Parallel Computing with Cloud and Cloud. 1-39.

[21] S. Kalid, A. Syed, A. Mohammad, and M. N. Halgamuge, "Big-Data NoSQL Databases: Comparison and Analysis of "Big-Table", "DynamoDB", and "Cassandra", IEEE 2nd International Conference on Big Data Analysis, Beijing, China, pp 89-93, 10-12 March 2017.

[22] V. Vargas, A. Syed, A. Mohammad, and M. N. Halgamuge, "Pentaho and Jaspersoft: A Comparative Study of Business Intelligence Open Source Tools Processing Big Data to Evaluate Performances", International Journal of Advanced Computer Science and Applications (IJACSA), Volume 7, Issue 10, pp 20-29, November 2016.

[23] A. A. R. Madushanki, M. N. Halgamuge, W. A. H. S. Wirasagoda, and A. Syed, "Adoption of the Internet of Things (IoT) in Agriculture and Smart Farming towards Urban Greening: A Survey", International Journal of Advanced Computer Science and Applications (IJACSA), Volume 10, No. 4, pp 11-28, April 2019.

[24] A. Singh, M. N. Halgamuge, R. Lakshmiganthan, "Impact of Different Data Types on Classifier Performance of Random Forest, Naïve Bayes, and k-Nearest Neighbors Algorithms", International Journal of Advanced Computer Science and Applications (IJACSA), Volume 8, No 12, pp 1-10, December 2017.

The New High-Performance Face Tracking System based on Detection-Tracking and Tracklet-Tracklet Association in Semi-Online Mode

Ngoc Q. Ly[1], Tan T. Nguyen[2], Tai C. Vong[3]
Faculty of Information Technology
VNUHCM-University of Science
Ho Chi Minh City, Viet Nam

Cuong V. Than[4]
AI Department
Axon Company
Seattle, USA

***Abstract*—Despite recent advances in multiple object tracking and pedestrian tracking, multiple-face tracking remains a challenging problem. In this work, the authors propose a framework to solve the problem in semi-online manner (the framework runs in real-time speed with two-second delay). The proposed framework consists of two stages: detection-tracking and tracklet-tracklet association. Detection-tracking stage is for creating short tracklets. Tracklet-tracklet association is for merging and assigning identifications to those tracklets. To the best of the authors' knowledge, the authors make contributions in three aspects: 1) the authors adopt a principle often used in online approaches as a part of the framework and introduce a tracklet-tracklet association stage to leverage future information; 2) the authors propose a motion affinity metric to compare trajectories of two tracklets; 3) the authors propose an efficient way to employ deep features in comparing tracklets of faces. The authors achieved 78.7% precision plot AUC, 68.1% success plot AUC on MobiFace dataset (test set). On OTB dataset, the authors achieved 78.2% and 72.5% precision plot AUC, 51.9% and 43.9% success plot AUC on normal and difficult face subsets, respectively. The average speed was maintained at around 44 FPS. In comparison to the state-of-the-art methods, the proposed framework's performance maintains high rankings in top 3 on two datasets while keeping the processing speed higher than the other methods in top 3.**

***Keywords*—Face tracking; face re-identification; detection-tracking; tracklet-tracklet association**

I. Introduction

While multiple object tracking has been receiving much attention from researchers all over the world, multiple-face tracking has received much less attention due to two main reasons: face tracking is a sub-problem of object tracking thus many works focus on the general problem, and there is a lack of encompassing multiple-face tracking datasets. Therefore, multiple-face tracking remains a challenging problem. Recent advances in the field of multiple pedestrian tracking can be used to solve the problem of multiple-face tracking. There are two main research directions for the problem: online and offline.

Offline approaches [1]–[6] treat the problem as a global optimization one and solve it once having received all the information of all frames of a video. These approaches basically revolve in three stages:

Stage 1: Apply detection algorithms over all frames of the video to get detected bounding boxes of individuals, which are treated as nodes of a graph.

Stage 2: Define a meaningful metric to measure the relationship between two nodes of the graph by employing visual, spatial and temporal information.

Stage 3: Optimize an objective function globally to get clustered the bounding boxes of individuals.

These approaches tend to use commonly known detectors to generate all detection boxes (stage 1). However, these methods are different from each other in defining relations between nodes (stage 2) and objective functions (stage 3). Berclaz et al. [1] propose to model all potential locations over time, find trajectories that produce the minimum cost and track interacting objects simultaneously by using intertwined flow and imposing linear flow constraints. Milan et al. [2] employ an energy function that considers physical constraints such as target dynamics, mutual exclusion, and track persistence. Tang et al. [4] propose to jointly cluster detections over space and time by partitioning the graph with attractive and repulsive terms. Cruz et al. [6] introduce two lifted edges for the tracking graph that add additional long-range information to the objective. The authors of [6] also employ human pose features extracted from a deep network for the detection-detection association. Solving the problem with no constraints of speed while having all the information beforehand, offline approaches often produce higher accuracy than online approaches summarized as follows.

Online approaches mainly focus on tracking by detection [7]–[15]. Basically, they employ three models: a state-of-the-art detection model to produce face detection bounding boxes, a standalone tracker [16]–[19] to produce face track bounding boxes, and a deep feature model [20]–[26] to extract representative features for matching. Combining detection and tracking methods help alleviate challenges when using stand-alone trackers such as sudden movements, blurring, pose variation. By adopting the detection-tracking framework, the problem of face tracking is then reduced to data association [27], [28] problem, that is to assign detection boxes to track boxes. Data association [27], [28] between detection boxes and track boxes then can be reduced to the bipartite matching problem (assume no two detection boxes in one frame belong to one individual, and so for track boxes) and can be efficiently solved by Hungarian algorithm [29]. Because bipartite matching algorithms find 1-1 matches, it is crucial to define a meaningful affinity metric, representing the relationship between two nodes, for good performance.

These online approaches can be simplified as follows:

Step 1: For each frame, run a detection model to get possible positions of faces in that frame (these results will be referred as detections). Then apply a deep feature model to extract features of these detections.

Step 2: Also, for that frame, run a tracker for each tracklet to get new possible positions from the previous position of each tracklet (these results will be referred as predictions). Then apply a deep feature model to extract features of these predictions.

Step 3: A defined metric is employed to relate detections with predictions. The metric consists of two parts: motion affinity and appearance affinity. Motion affinity is measured by the intersection over union (or Mahalanobis distance) of detections and predictions. Appearance affinity is measured by Euclidean (or cosine) distance between features of detections and features of predictions (or possibly of tracklets).

Step 4: After three steps above, the result is an affinity matrix (N detections x M predictions). Apply a bipartite matching algorithm to associate new detections with predictions. Unassigned detections are treated as new individuals while assigned detections are used to update tracklets.

Step 5: Repeat steps 1-4 consecutively for frames of a video.

There are some disadvantages to these online approaches.

Disadvantage 1: At the i-th frame, new detections must be assigned identifications at that frame. This means the information in the future cannot taken advantage of.

Disadvantage 2: To decide whether a new detection belongs to a known identity or is a new identity, the similarity matrix (computed by motion and appearance affinity) is used. To have the number of tracklets for one individual as low as possible, the threshold must be lowered. However, doing that way, the possibility of one track containing many individuals is high.

Disadvantage 3: Because detection-tracking method must run detection model and tracking algorithm for each frame to get new detections and new predictions, then run deep feature model (models used for feature extraction are computationally expensive) for new detections and new predictions, these models must be lightweight to run in real-time. This can lead to low accuracy in these models and causes errors for the whole framework.

Disadvantage 4: Because these approaches compare detections with predictions, they fail to employ very potential information that can be taken advantage of when comparing tracks to tracks. That is the fact that two temporal-overlapped tracks cannot belong to the same individual.

To resolve the issues stated above, the authors propose a semi-online framework for the multi-face tracking problem. The framework consists of two stages: detection-tracking stage and tracklet-tracklet association stage. For the detection-tracking stage, the authors employ the same principle as in online approaches with a modification: the authors use two complementary trackers (Kalman filter as a motion tracker and KCF (Kernelized Correlation Filter) as a visual tracker) to improve accuracy. For the tracklet-tracklet association, inspired by offline approaches, the authors treat each tracklet as a node of a graph and optimize the problem of assigning identifications globally. In this stage, the authors also introduce an efficient metric to compare two tracklets so that the framework can run with high speed.

The rest of this paper is organized as follows. In Related Works, the authors begin to cover current state-of-the-art methods for multiple-face tracking in two modes: offline and online. In Materials and Methods, the authors then turn to the proposed approach which is inspired by principles used in both offline and online multiple-face tracking. In this section, the authors illustrate the overview and detailed stages of the proposed framework. The authors conclude this section with contributions to literature. In Results and Discussions, the authors describe experiments and datasets, report experimental results, and discuss some implications. The final section concludes the proposed approach and considers ways to further improve multiple-face tracking.

II. Related Works

A. *Offline Tracking*

State-of-the-art methods for multi-face offline tracking are [30]–[32]. These approaches can be reduced to two main stages: tracklet creation (tracking-by-detection) and tracklet association. In [30], Zhang et al. first divide the video into many non-overlapping shots – music or film videos often contain many shots in different scenes. For each shot, the framework employs the tracking-by-detection paradigm to generate tracklets and merge those tracklets into groups by temporal, kinematic (motion, size) and appearance (deep feature) information. Then, Zhang et al. link tracklets across shots/scenes by treating each tracklet as a point, the appearance similarity between two tracklets as edge and applying the Hierarchical clustering algorithm to assign tracklets into groups. To increase the accuracy of the tracklet linking step, a discriminative feature extractor is needed. The authors of [30] introduce Learning Adaptive Discriminative Features whereby a deep extractor will be finetuned online based on samples from the video. Jin et al. [31] improve the performance of the mentioned method by using a more powerful detector (Faster R-CNN) in the tracking-by-detection stage and a more sophisticated tracklet association schedule. Lin et al. [32] push it further by applying body parts detector and introduce a co-occurrence model to generate longer tracklets when faces are out of camera (but body not) or detector cannot capture faces. Besides, the work also introduces a refinement scheme for tracklet association based on Gaussian Process.

B. Online Tracking

1) Hand-crafted features: One of the attempts to solve the multi-face online tracking problem that yield good results is [33]. In this work, Comaschi et al. adopt the tracking-by-detection mechanism for the pipeline (Fig. 1). Because of the frontal characteristics of the dataset being used, the work employs a Haar-like cascade face detector [34] to attain computational efficiency. In any tracking problem, the ability to learn appearance change and predict future states of objects is crucial for the model. Thus, the work introduces a structured SVM tracker that stores previous patterns and positions of an object and can predict the new state of an object based on current spatial and visual information. The tracker is updated online based on both track prediction and detection. In the data association step, this work applies Hungarian algorithm for the cost matrix computed by the intersection over union of detection boxes and track boxes.

Similar to the above work, Lan et al. [35] also adopt tracking-by-detection mechanism but with a more sophisticated tracker update routine. Naiel et al. [36] try to decrease the false negative rate (miss detection caused by a simple detector) of the previous pipeline without reducing speed. In this work, Naiel et al. adopt an advancement of [34] and a color-assisted tracker as detect and track components respectively (Fig. 2). The novelty of this work lies in the combined framework. Instead of running a detector for every frame like previous work, Naiel et al. propose a trigger mechanism so that the detector only need to run on some specific frames. Specifically, the detector is only triggered after a fixed interval (N frames) or earlier, when there is any tracking fail. The authors compare the histogram of the new track box with histograms of previous track boxes. If there is any large discrepancy, the track fail will trigger detection.

Similarly, the authors of [37] adopt the idea of sparse detection, modifies Viola-Jones detector in conjunction with a variant of optical flow to create a combined detection-tracking model.

2) Deep features: Recently, many works [38]–[42] integrate deep feature extractors into the tracking framework. Of those works, Chen et al. [38] adopt the sparse detection mechanism as described above and use KLT tracker [43] for the tracking-by-detection stage. In the data association step between detection boxes and track boxes, deep feature vectors are used as visual information in addition to spatial information.

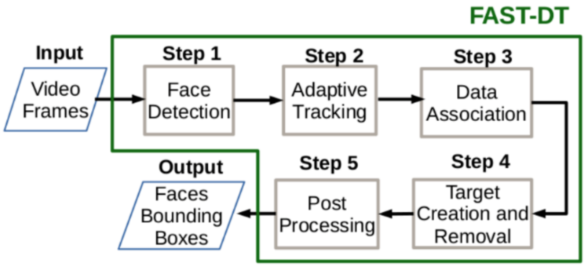


Fig. 1. Multi-Face Detection and Tracking Framework [33].

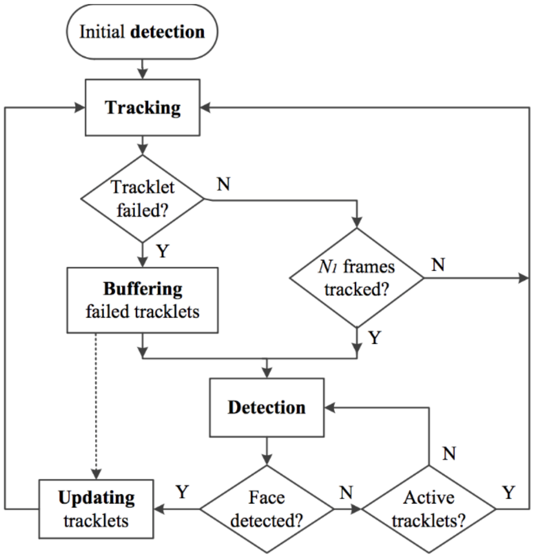


Fig. 2. Multi-Face Tracking Detection and Tracking Flow [36].

III. Method

A. Overview

1) Semi-online tracking: Aiming for practical usage and from the analysis of the online detection-tracking approaches, the authors propose a new approach in semi-online manner by introducing the tracklet-tracklet association stage (Fig. 3).

After getting the detections of a frame, the authors should match it with tracklets up until the previous frame to determine identifications for new detections. To achieve this criterion, using a deep feature extractor is a heavy waste. The authors propose a way to lighten the process while keeping the accuracy as high as possible. First, the authors use a light feature LBPH (Local Binary Pattern Histogram) extractor in the detection-tracking stage (Fig. 5) for efficient computation and combine it with information from a tracking method (Kalman filter) to reduce the errors as much as possible in creating short tracklets (the authors have not yet assigned identifications for those tracklets). Then the authors observe that consecutive face boxes of one tracklet are nearly the same, thus in the tracklet-tracklet association stage (Fig. 7), the authors introduce a compression method to get representatives of a tracklet and apply a deep feature extractor on these representatives instead of all boxes. The authors then link short tracklets into long tracklets by using those features as appearance information. In the linking step, the authors also introduce a new method for motion similarity between two tracklets. The tracklet-tracklet association stage resolved much problems stated above: the future information of frames sequences is well manipulated; the computational complexity is cut off from deep feature comparison by applying the new compression method.

Detection-Tracking stage: The main role of this stage is to extract the track information of targets in a frame using detecting and tracking methods. Technically, the detection-tracking stage processes frame-by-frame for every mini-batch interval (64 frames) and yields a list of tracklets. The process is illustrated in Fig. 4.

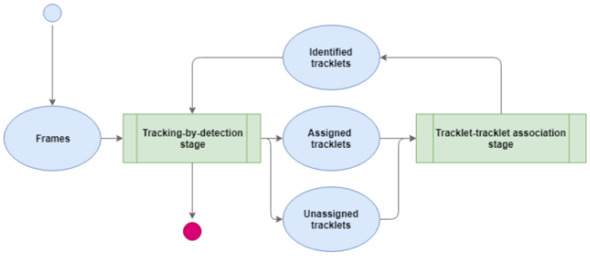


Fig. 3. Our Proposed Method. The Extra Tracklet-Tracklet Association is Introduced to Improve Accuracy by using more Information and Lighten the Process before.

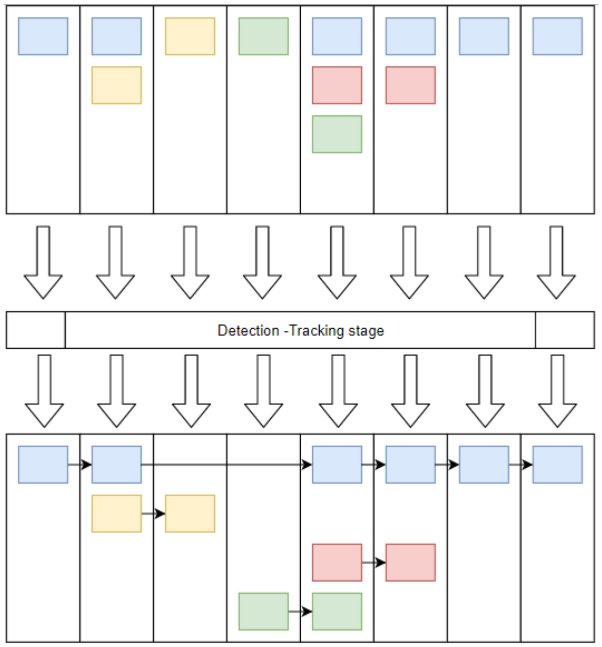


Fig. 4. Detection-Tracking Stage (Frame by Frame). Columns are Consecutive Frames; each Box is a Tracked Box in each Frame; the Arrows show how a Tracklet is Formed; Each identity is Marked by different Colors in Each Box.

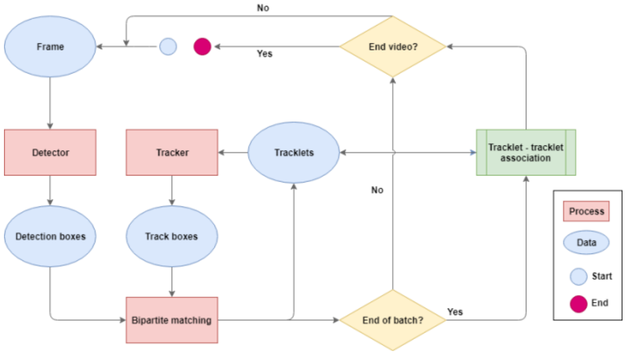


Fig. 5. Our Detection – Tracking flow Diagram.

The end-to-end framework consists of two stages:

Tracklet-tracklet association stage: At the end of each mini-batch process, the list of tracklets is passed to this stage. The main role of this stage is to correct false positives of the previous stage and connect related tracklet to create long tracklets and then assign identifications to these new tracklets. The process is shown in Fig. 6.

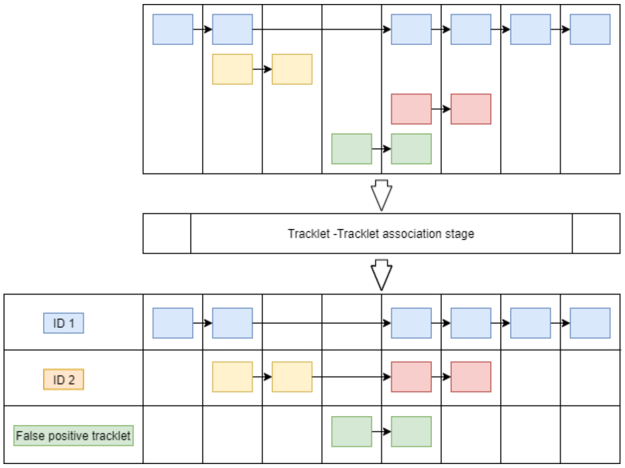


Fig. 6. Tracklet-Tracklet Association Stage. from Tracklets Formed before, the Identities will be Determined in this Stage.

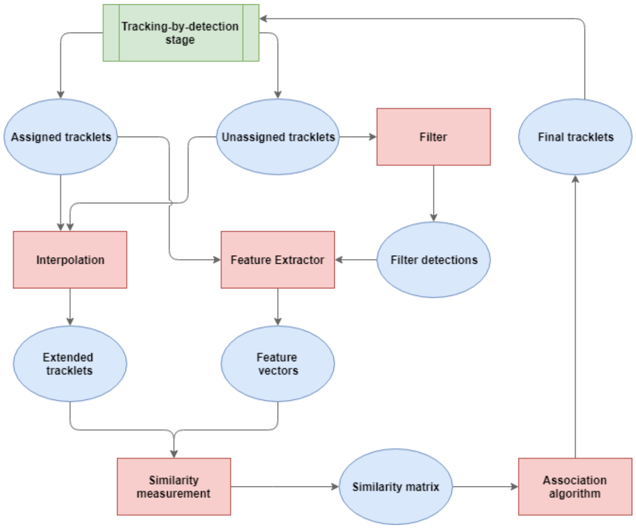


Fig. 7. Our Tracklet-Tracklet Association flow Diagram.

The proposed framework returns results after the tracklet-tracklet association stage. For instance, it returns results of frames 1-st to 64-th after seeing the information of frame 64-th. This induces a delay of over 2 seconds (64 frames ~ 2 seconds in normal 30fps videos). The details of the proposed framework are explained follow.

2) Computational complexity: The proposed framework can process video streaming in real-time. The speed can reach around 60fps, which is greater or equal the frequency of common videos (from 30 to 60fps).

3) Detection-tracking stage: The authors leverage known detection-tracking approaches with some modifications to speed up the stage without sacrificing much performance and introduce a new stage to improve the performance. The authors also implemented a framework: the detection-tracking stage combining S3FD face detector to produce detection boxes, LBPHs feature extractor to extract the global features, Kalman Filter tracker to produce tracking boxes, then Hungarian algorithms for matching the corresponding boxes to create tracklets.

4) Tracklet-tracklet association stage: The tracklet - tracklet association stage uses the motion information simulated by the spline interpolation and appearance information from FaceNet deep feature extractor to drop the false positives and match the suitable tracklets to accurately assign the ids for targets.

B. Detection – Tracking Stage

1) Goal: In this stage, all the detection boxes of all frames in a batch will be grouped into short tracklets with the help of a single object tracking method.

2) Principle: Combining a single tracker and a detector helps a lot in overcoming the limitation of each single method. Using single trackers [16]–[19] to track faces in the wild situation is hard due to occlusion, illumination change, pose variation, sudden movement, etc. These issues can lead to track losses, inaccurate boxes (boxes that capture part of the face), incorrect boxes (boxes that capture the face of another individual). Moreover, using only a detector faces the appearance feature confusion if there are faces of different individuals with high appearance similarity.

The authors observe that detection models yield neater boxes than single trackers so using detection boxes as new information for updating single trackers is reasonable.

3) Method: In this stage, a detection model is used to generate possible bounding boxes of faces in a frame. During that time, a tracker is also used to predict a new possible bounding boxes positions from previous frames. Our detection-tracking algorithm will try to fuse these detection results with track results in order to better enhance the output, create more reliable tracklets.

At each frame, after running the detection and tracking process, the authors get a list of (N) detection boxes and (M) track boxes. The track boxes are the spatial predictions of bounding boxes from previous tracklets, while the detection boxes are the bounding boxes of faces that existed in that frame. Those faces may be the old faces from the previous frames, but they may also be the new faces that only exist from that frame. The main purpose of the detection-tracking algorithm is to define a meaningful affinity matrix (N x M) so that it can reflect the relationships between those detection boxes and track boxes.

Two features that are commonly leveraged are motion and appearance:

Motion affinity between a detection box and a track box is defined by the intersection over union (IoU) of them.

Appearance affinity between a detection box and a track box is defined by cosine affinity between LBPH features of them.

Those two features are used because for a pair of detection box and track box to be matched, two boxes should be close to each other with similar size and visual feature.

The authors define a gating unit for each affinity in order to filter out less likely matches. Because of our intention that if a detection box and a track box are considered a possible match, they must satisfy motion affinity alone and appearance affinity alone first.

As explained, the authors want both metrics to be high to treat a pair of detection box and track box a likely match; thus, if both affinity metrics pass the threshold then the final affinity is the multiplicative result of motion and appearance affinity, otherwise is zero.

$$Match(i,j) = \begin{cases} s_m(i,j).s_{\mathrm{a}}(i,j) \\ if\ s_m(i,j) > \gamma_{\mathrm{M}}\ and\ s_{\mathrm{a}}(i,j)\ \ > \gamma_{\mathrm{A}} \\ 0 \qquad else \end{cases} \tag{1}$$

where,

$\mathrm{s_a}(i,j)$ describes the appearance similarity distance between bounding boxes i and j, its range is from 0 to 1.

$\mathrm{s_m}(i,j)$ describes the space similarity distance between bounding boxes i and j, its range is from 0 to 1.

γ_{M} is the threshold for space similarity distance determined by heuristic (the authors reason that detection box and track box should be near to be of one individual, so the authors set this value to 0.3).

γ_{A} is the threshold for appearance similarity distance determined by heuristic (the purpose of this stage is to create short tracklets, the authors use a high threshold to prevent wrong matches, specifically 0.9).

$Match(i,j)$ will be used to determine if a detection box and a track box is a possible match. It only has value if both motion and appearance metrics are over their thresholds. If one of the metrics is lower than its respective threshold, $Match(i,j)$ is set to 0. The thresholds for $Match(i,j)$ are determined through experiments (value search).

C. Tracklet-Tracklet Association

1) Goal: Short tracklets from the detection-tracking stage are passed to this stage. The authors will group short tracklets into long tracklets and assign identifications for them. After this stage, the boxes in each frame will be marked with identifications and ready to deliver to the result stream.

2) Principle: The objective of face tracking is that for everyone existed in a video, the framework should output as few as possible the number of tracklets for that individual without wrongly including other faces of other individuals. This leads to the tradeoff mentioned in Section I. The authors tackle this with two principles:

Make sure the possibility of wrongly matching is as low as possible by using tight constraints (high affinity thresholds).

Adopt efficient motion and appearance affinity metrics between tracklets (different from track-detection) to group tracklets into identities based on a community discovery algorithm in this stage.

3) Method: After each batch processing the detection-tracking stage, the authors have a list of unknown-id tracklets that are needed to be assigned identifications in this stage. The authors also have a list of known-id tracklets in the past (previous batches). Our job is now trying to assign identifications to unknown-id tracklets.

The authors formulate the assignment puzzle as an optimization problem. Each tracklet is treated as a node of a graph. The edge of two nodes indicates the affinity between the two. The authors then apply a clustering algorithm, in this situation, Leiden algorithm [28] on this graph in order to partition it into subgraphs – groups, each containing tracklets - nodes of the same individual. The authors put constraints so that each subgraph will not contain two known-id tracklets or two temporally overlapped tracklets. One of the essential parts of this stage is defining a meaningful metric representing the edge of two nodes. To do that, the authors adopt the complementary nature of motion and appearance.

a) Motion distance: For motion, the authors introduce a trajectory difference metric. Given two tracklets (t(i), t(j)), it is safe to assume that t(i) predate t(j) and there is no temporal overlap between two tracklets. From the boxes of t(i), the authors extrapolate forward to get the possible boxes in the future relative to t(i). From the boxes of t(j), the authors extrapolate backward to get the possible boxes in the past relative to (t(j). For extrapolation, the authors assume that face movement can be modeled as a polynomial function and apply spline extrapolation. The authors ran model selection to determine the degree of movement and found that 1-degree spline performs best. Now the extrapolated parts of the two overlap temporally, the authors have a pair of overlapped extrapolated boxes in the same frame f(k). The authors now calculate a spatial distance between two boxes using two centers and a diagonal distance between two boxes according to their diagonals. The authors introduce a weight parameter to fuse the two distances into one unified box-box distance.

The box-box distance at frame k can be formulated in the following equation:

$$d_{M,k} = \lambda . d_{S,k} + (1 - \lambda) . d_{D,k} \gamma \tag{2}$$

In that,

$d_{S,k}$ is the Euclidean distance between two centers of two boxes.

$d_{D,k}$ is the diagonal distance between two boxes calculated by the difference in length between two diagonals.

λ is the weight parameter to fuse above distances into one unified distance (the authors search from 0 to 1 with 0.1 interval and choose 0.4 to maximize area under the curve of success plot).

$d_{M,k}$ is the box-box distance at frame k the authors are going to obtain.

Then the trajectory distance is the average of pair distances:

$$d_M = \frac{1}{n-m+1} \sum_{k=m}^{n} d_{M,k} \tag{3}$$

where,

$k = m \rightarrow n$ are overlapped frame indices.

$d_{M,k}$ is the box-box distance at frame k.

d_M is the trajectory distance, the average box-box distance over $m - n + 1$ frames.

b) Appearance distance: For appearance, the authors use average Euclidean distance between two feature sets of two tracklets. For each box of a tracklet, the authors have a respective LBPHs feature (referred to as light feature) extracted from the detection-tracking stage. Assume t(i) have N light feature vectors and t(j) have M light feature vectors, one straightforward method is to compute N*M Euclidean distances and use the average as the distance between two tracklets. However, the task is to distinguish between human faces, LBPHs feature is not discriminative enough for this task that requires fine-grained features. Besides, deep neural networks have outperformed hand-crafted methods on many visual tasks that require fine-grained features. Thus, the authors employ a deep feature extractor (Facenet) [20] for this task. Specifically, the authors deploy the pretrained model and feedforward to extract features.

However, deep feature extractors are computationally expensive and if the authors compute deep features for all boxes of a tracklet the framework would not run in real-time. Moreover, temporally adjacent boxes often contain similar information, so it would be redundant to compute all the deep features. The authors introduce our compression method to lower the number of boxes needed to be passed through a deep feature extractor using already computed light features.

Given a list of light feature vectors of a tracklet, the authors apply a clustering algorithm on these light feature vectors and pick out centroids, i.e. $N_{compressed}$ boxes, for deep feature extraction. Only centroids are then passed to the deep feature extractor to extract 128-dimensional vectors. This way the authors save a lot of time computing deep features while keeping the diversity of a tracklet. The authors then use average Euclidean distance between two deep feature sets of two tracklets as tracklet - tracklet appearance distance:

$$d_A = \frac{1}{N_{compressed}} . \frac{1}{M_{compressed}} \times \sum_{n1}^{N_{compressed}} \sum_{m1}^{M_{compressed}} Euclid\big(f(n), f(m)\big) \tag{4}$$

In that,

$M_{compressed}$ is the number of filtered boxes of the first track for deep feature extraction.

$N_{compressed}$ is the number of filtered boxes of the second track for deep feature extraction.

d_A is our tracklet – tracklet appearance distance, calculated as the average Euclidean distance between two deep feature sets of two tracklets.

$f(n)$ is the feature extracted from the n-th box of $N_{compressed}$ boxes.

$f(m)$ is the feature extracted from the m-th box of $M_{compressed}$ boxes.

c) Fusing results: A weighted sum of appearance and motion affinities is the affinity between two tracklets (used as the weight of the edge between two nodes). The authors fuse two affinities by taking the addition rather than multiplication as used in the detection-tracking stage because motion affinity is not reliable enough in case of long-term occlusion or camera shake. Thus, the authors set the weight for motion affinity low so that it plays as extra information.

$$d_{AM}(i,j) = \lambda . d_M(i,j) + (1 - \lambda) . d_A(i,j) \qquad (5)$$

Where

$d_M(i,j)$ is the motion dissimilarity distance, calculated as explained.

$d_A(i,j)$is the appearance dissimilarity distance, calculated as explained.

λ is the weight parameter to adjust the importance of each distance. This value is determined through experiments (the authors search from 0 to 1 with 0.1 interval and choose 0.3 to maximize area under the curve for success plot).

$d_{AM}(i,j)$ is the dissimilarity distance of tracklet i and j.

D. Contributions

This proposed approach tackles challenges related to online approach above:

- Instead of computing deep features for all faces of one tracklet as online approaches do, the authors leverage light features (LBPHs) in the context of tracklet to efficiently compute deep features (extracted by deep network) without compromising representative power. In fact, the compressing method produces a more accurate representation for a tracklet thanks to diversity and high detection quality (high-score detected boxes).

- Using this framework, the authors can tighten the constraints in the tracking-by-detection stage so that the possibility of wrongly matching is low. Though having many tracklets after the tracking-by-detection stage, these tracklets will be grouped in the tracklet-tracklet association stage.

- The authors do not have to assign identifications to new detections right away in the detection-tracking stage but leave it to the tracklet-tracklet association stage. This way the authors can filter out false positives efficiently in the pre-processing step.

- The identification assignment step is tracklet-based; thus, the authors can take advantage of temporal information of tracklets (co-extant tracklets belong to different individuals).

- The authors also propose the trajectory difference metric to account for motion in tracklet-tracklet comparison.

In application, dataset is limited so using a pre-trained model and finetuning on small dataset is a reasonable choice. In this work, the authors show that simply adopting deep features (extracted by Facenet) and employ Euclidean (or cosine) metric is not discriminative enough in reference to real-life data. Therefore, the authors propose to apply Logistic discriminant metric learning so that the new embedding space for real-life data is more discriminative.

The authors speculate that other regions of person, besides the face, also contain discriminating features. The authors tried to employ some color-based feature (color name) and texture-based feature (LOMO) but the results were not comparable, thus leaving this part for future work.

IV. Results and Discussions

Our experiments are conducted by python on the hardware GTX 1080 GPU, Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz, 16GB RAM, while the MobiFace paper [44] used a desktop machine with Intel i9-7900X CPU (3.30GHz) and one GTX 1080 Ti GPU. Therefore, it's fair to compare the speed of our method versus other methods on MobiFace. For OTB dataset [45], RFTD method [46] used a setup with Intel Core i7 with 3.07GHz clock with no GPU and CXT and SCM used similar computational power, so the authors only compare the performance of our method versus other methods in terms of accuracy.

A. The Purpose of Experiments on MobiFace and OTB Datasets

In order to prove the efficiency of our tracking framework, the authors conducted two comparisons:

Comparing single trackers with tracking-by-detection approaches through results from MobiFace Dataset. The purpose is to prove that integrate the detection method will enhance the result more than using a single tracker.

Comparing tracking-by-detection approaches with our approach through results from OTB Dataset. The purpose is to prove that using the light feature to process in the tracking-by-detection stage and using the deep feature in the tracklet - tracklet association stage in conjunction with motion affinity is a significant improvement.

1) Experiments on MobiFace dataset

a) About the dataset: MobiFace dataset [44] is the first dataset for single face tracking in mobile situations. Due to the lack of engrossing face tracking datasets before MobiFace, the performance of pioneer face trackers was reported on a few videos or on small subsets of the OTB dataset, and the comparison between approaches was limited. The introduced dataset provides a unified benchmark with different attributes for future development in this field. Some samples of the dataset are illustrated in Fig. 8.

The authors collected 80 unedited live-streaming mobile videos captured by 70 different smartphone users in fully unconstrained environments and manually labeled over 95.000 bounding boxes on all frames. In order to cover typical usage of mobile device camera, the authors fetched videos from YouTube mobile live-streaming channels. Most of the videos are captured and uploaded under fully unconstrained environments without any extra video editing or visual effects. 6021 videos were collected and discarded under strict criteria that the target faces should appear at least in 10% of the video frames, and the target faces should not always stay still to serve the purpose of visual tracking. Besides the common 8 attributes in object tracking datasets, the authors proposed six additional attributes commonly seen in mobile situations.

The authors also fine-tuned and improved a handful of state-of-the-art trackers and perform evaluations on the dataset. Through comparing with those results, the authors can evaluate the efficiency of our method.

b) Setup the experiments: Note that MobiFace dataset is designed for supervised trackers - an initial box of a targeted face is specified in the first frame. However, our method is designed to work in an unsupervised way (the authors do not need initial boxes) and can track multiple targets at a time. In order to adapt to the dataset, the authors must reduce the system to fit with the protocol of the dataset. Specifically, in the first frame of each video, the authors compare the detected result of our system with the initial box provided by the dataset to specify the targeted face and then return track results of that target only.

The video is only stored in YouTube so from the time the authors access it, the authors are unable to collect all videos from the dataset because some has been deleted by the owners.

The authors consider the three metrics proposed in the dataset: normalized precision, success rate, frames per second. As most of the metrics are in plot form, the authors will explain the way to extract an important metric from the plot, the area under the curve (AUC). With N is the number of thresholds used to draw the plot and $n = 1, 2, 3, \ldots, N$. The curve was drawn from points with coordinate (t_n, f_n), t_n is the threshold value at that point and f_n is the evaluated value of our algorithm at that threshold, i.e. location error of precision plot, overlap score of success plot. The AUC is then calculated by

$$AUC = \sum_n (t_n - t_{n-1})\, f_n \qquad (6)$$

Normalised precision plot: Precision plot is a widely used evaluation metric for the tracking field. The precision is described as the location error, which is the Euclidean distance between the center location of the tracked face and the ground truth bounding box. This metric reflects how far the tracker has drifted from the targeted face. However, as the videos differ greatly in resolution, the authors of [44] adopt the recently proposed normalised precision value. The size of the frame is used for the normalisation, and the authors of [44] rank the trackers based on the area under the curve (AUC) for normalised precision value between 0 and 0.5.

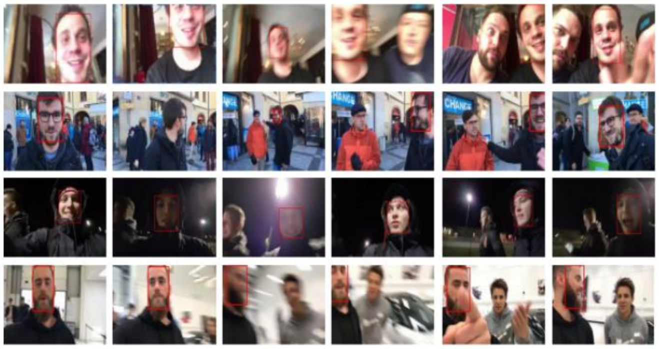


Fig. 8. Some Example Frame from the MobiFace Dataset [44]. Red ground Truth Bounding Boxes are Annotated by the Authors.

Success plot: Overlap score is also another commonly used metric in the tracking field. Given a ground truth bounding box r_{gt} of the target, the predicted bounding box of our algorithm is r_p. Then the overlap score can be computed by the intersection over union (IoU) of those two boxes as $S = \frac{r_{gt} \cap r_p}{r_{gt} \cup r_p}$, where the $\cap$ and $\cup$ represent the intersection and union of two rectangles, respectively. The success plot reflects the percentage of frames in which the intersection over union (IoU) of the predicted and ground truth bounding box is greater than a given threshold. Usually, the average success rate at 0.5 threshold is enough for evaluation. In addition, the area under the curve (AUC), which is the accumulated success rate can also be used for measurement. The authors can use those metrics interchangeably to summarize the performance.

Frames Per Second (FPS): the average speed of the evaluated tracker running across all the sequences. The initialization time is not considered. Because of the applicability concern, a mobile face tracker must be able to run at high speed (either on CPU or GPU) to allow maximum potential migration to actual mobile devices. Due to the lack of implementation of competitive trackers on mobile platforms, the authors can only use the FPS measured on the desktop environment, which indicate the relative efficiency of the trackers for evaluating and comparing.

c) Experiment results: Evaluation metrics of our method and state-of-the-art methods are illustrated in Fig. 9 and a detailed comparison is shown in Table I.

TABLE. I. A Detailed Comparison between our Method and Mobiface Evaluated Results

Tracker	Normalised Precision plot (AUC)	Success plot (AUC)	FPS
MDNet-MBF+R	**0.800**	**0.601**	1.79
MetaMDNet-MBF+R	0.767	**0.571**	1.03
MetaMDNet-YTF+R	0.744	0.566	1.06
MDNet-MBF	**0.772**	0.549	1.58
SiamFC-MBF+R	0.758	0.526	**53.14**
SiamFC-MBF	0.750	0.521	**81.54**
Proposed framework	**0.787**	**0.681**	**44.38**[a]

[a] The authors profile the program and exclude reading image from disk time and writing image to disk time before calculating speed (details are in test.profile file in our source code).

d) Discussion: Because our approach is targeted for the multi-face tracking field. In order to make it work with the dataset, the authors run the framework over the dataset and get all tracks of targets in the video, then according to the initialized ground truth box, the authors define the target and return the target track results only. Because the dataset is from unconstrained environments with many existing faces, it is a noticeable effort of our tracker to avoid mistakes between tracklets and output the correct results.

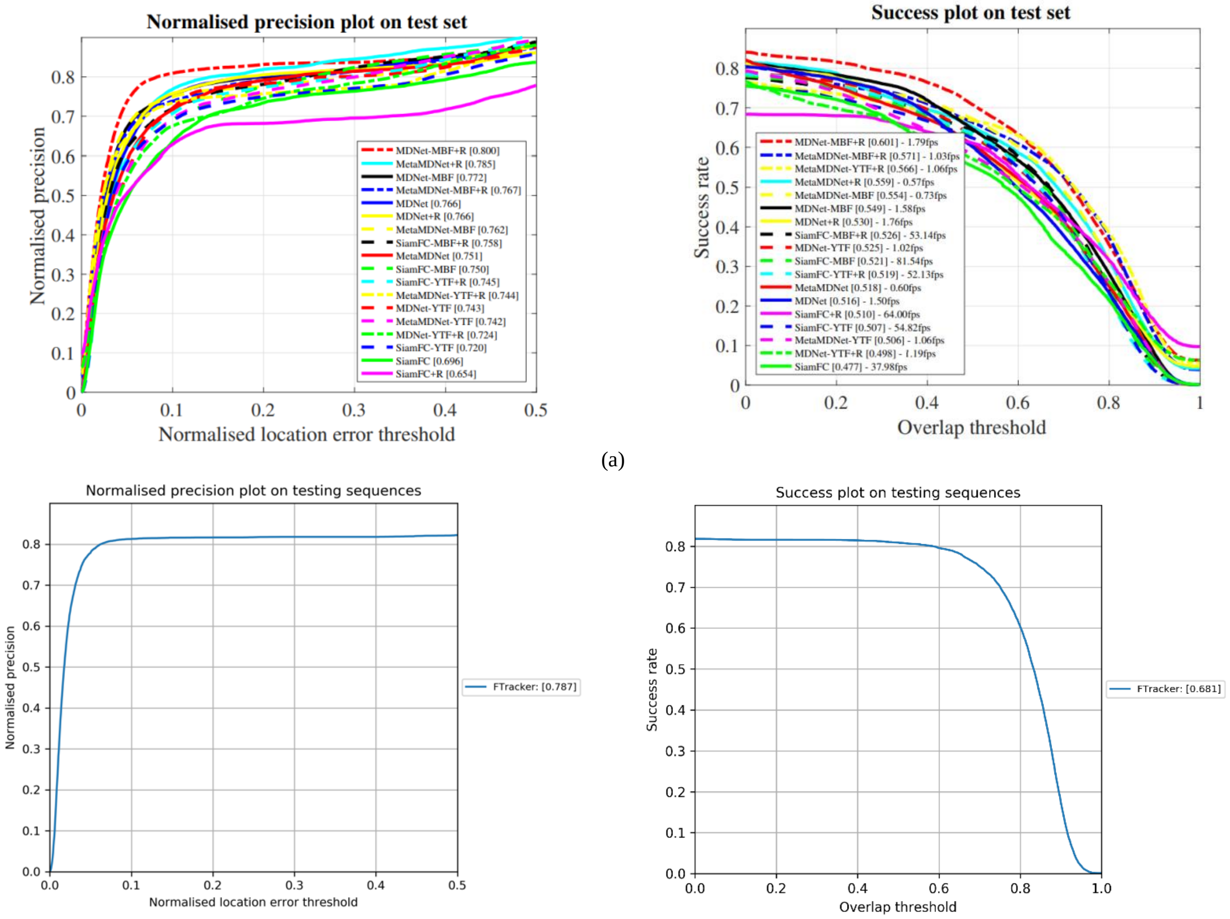
As shown in the above plots, our method has an advantage in the success plot, but not the precision plot. Precision is affected by the Euclidean distance between the center of a ground truth bounding box and the center of a tracked box. Because high normalised error still treats a tracked box that drifts out of a face (high Euclidean distance between two centers) as a true prediction, trackers that still maintain a track when the box drifts out of a face perform better with high normalized error. In the proposed framework, when the tracked box drifts out of a face, the algorithm terminates the tracklet instantly; therefore, with high normalised error, our tracker performs the same as with low normalised error while other trackers yield noticeably different results with different normalised errors.

The success plot might be more practical for applications that require high IoU between prediction boxes and ground truth boxes. The success plots of trackers evaluated in MobiFace dataset start very high, but the slope is very steep. Starting from above 0.8 success rate for threshold 0, to threshold 0.5, they drop to below 0.7 success rate. The steep slope indicates predicted boxes of those trackers are not always aligned with ground truth boxes. Our starting point is somewhere below 0.8 success rate but maintains the success rate over the overlap threshold change. At threshold 0.5, our approach still has a high success rate, above 0.7, indicating our boxes is closely aligned with ground truth boxes. At 0.5 threshold, the predicted boxes cover most of the track target and can be well used in application. Besides, as the main target of ours is for practical usages, a good success plot and success rate at 0.5 threshold - while keeping the speed - are acceptable.

2) Experiments on OTB (Object Tracking Benchmark) dataset

a) About the dataset: OTB Dataset [45] is one of the most famous datasets specifically used for benchmarking the object trackers since its appearance. The authors worked to collect and annotate most of the common tracking sequences from different datasets. They also classified those sequences into multiple categories by challenges as in Table II and selected 50 difficult and representative ones in the TB-50 dataset for an in-depth analysis. The full dataset contains more sequences of human (36 body and 26 face/head videos) than other categories because human target objects have the most practical usages, some samples of the dataset is illustrated in Fig. 10.

Before the introduction of MobiFace dataset, face tracking methods could only be evaluated on small self-collected datasets or a subset of OTB dataset. The whole dataset is designed for the object tracking algorithms, but the authors selectively pick out the sequences with faces to conduct experiments and compare with those methods mentioned before. The chosen face subset is described in Table III, the top 10 sequences are referred to as the difficult set and top 15 is the normal set [46].



(a)

(b)

Fig. 9. Evaluation Results of Trackers on MobiFace Test Set: (a) Results from MobiFace Paper [44], (b) Results on our Method.

TABLE. II. ANNOTATED SEQUENCE ATTRIBUTES WITH THE THRESHOLD VALUES IN THE PERFORMANCE EVALUATION FROM OTB DATASET [45]

Attribute	Description
IV	Illumination Variation - The illumination in the target region is significantly changed
SV	Scale Variation - The ratio of the bounding boxes of the first frame and the current frame is out of range. $\left[\frac{1}{t_s}, t_s\right], t_s > 1 (t_s = 2)$
OCC	Occlusion - The target is partially or fully occluded.
DEF	Deformation - Non-rigid object deformation.
MB	Motion Blur - The target region is blurred due to the motion of the target or the camera.
FM	Fast Motion - The motion of the ground truth is larger than t_m pixels ($t_m = 20$)
IPR	In-Plane Rotation - The target rotates in the image plane.
OPR	Out-of-Plane Rotation - The target rotates out of the image plane
OV	Out-of-View - Some portion of the target leaves the view
BC	Background Clutters - The background near the target has similar color or texture as the target
LR	Low Resolution - The number of pixels inside the ground-truth bounding box is less than t_r ($t_r = 400$)

TABLE. III. ANNOTATED SEQUENCE ATTRIBUTES WITH THE THRESHOLD VALUES IN THE PERFORMANCE EVALUATION FROM OTB DATASET [45]

#	Sequence	Challenge
1	**Soccer**	IV, SV, OCC, MB, FM, IPR, OPR, BC
2	**Freeman4**	SV, OCC, IPR, OPR
3	**Freeman1**	SV, IPR, OPR
4	**FleetFace**	SV, DEF, MB, FM, IPR, OPR
5	**Freeman3**	SV, IPR, OPR
6	**Girl**	SV, OCC, IPR, OPR
7	**Jumping**	MB, FM
8	**Trellis**	IV, SV, IPR, OPR, BC
9	**David**	IV, SV, OCC, DEF, MB, IPR, OPR
10	**Boy**	SV, MB, FM, IPR, OPR
11	**FaceOcc2**	IV, OCC, IPR, OPR
12	**Dudek**	SV, OCC, DEF, FM, IPR, OPR, OV, BC
13	**David2**	IPR, OPR
14	**Mhyang**	IV, DEF, OPR, BC
15	**FaceOcc1**	OCC

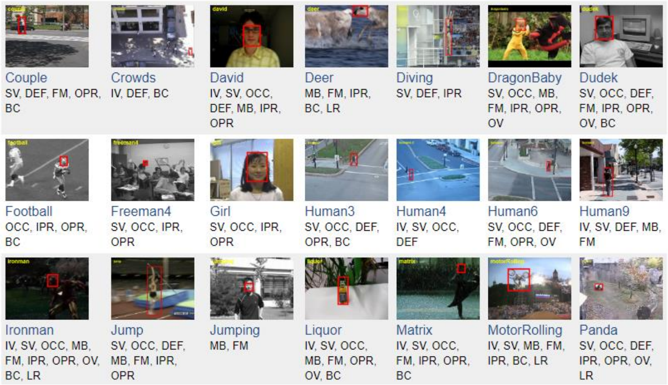


Fig. 10. Some Example Sequences from the OTB Dataset [45].

However, the dataset is also designed for the single object tracker. So, evaluation on this dataset also cannot reflect all the potential power of our system, but the authors can use that result to relatively compare with previous trackers in order to verify the power of the proposed framework.

b) Set up the experiments: Because the authors of MobiFace dataset inherit a lot of legacy from OTB dataset, in general, the setup stage and evaluation stage for OTB Dataset are the same as the MobiFace dataset.

c) Experimental results: Evaluation metrics of our method and state-of-the-art methods are illustrated in Fig. 11, Fig. 12, and a detailed comparison is shown in Table IV and Table V.

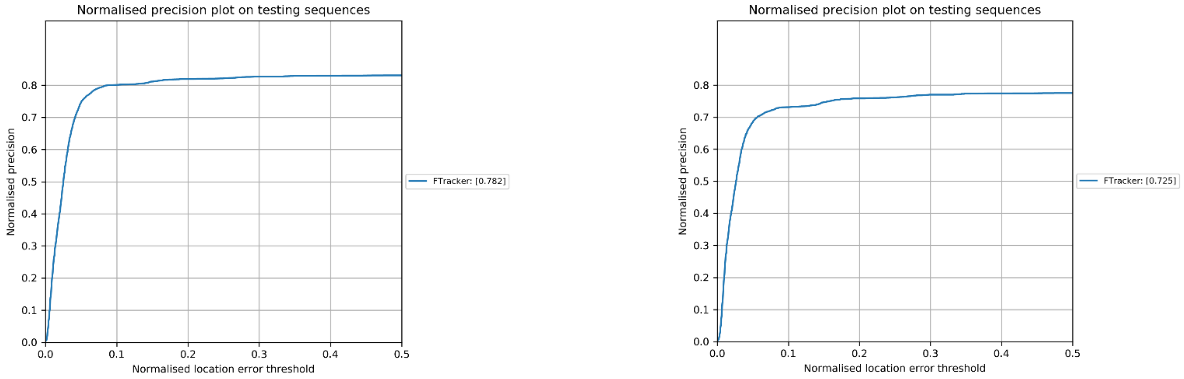
d) Discussion: The precision plots in Fig. 11 are good. The overall results are quite good, and the slope is shallow as predicted after witnessing above experiments. However, the authors have no data from other works to have an in-depth comparison.

TABLE. IV. TOP TRACKER COMPARISON ON OTB DATASET FACE SUBSET (NORMAL SET). EVALUATED RESULTS ARE FROM RFTD PAPER [46]

Face Tracker	Success Plot AUC	Success plot Threshold (0.5)
RFTD	55.2	**71.3**
Struck	**55.9**	67.6
SCM	**58.3**	**72.6**
ASLA	53.8	62.9
CSK	48.0	56.8
L1APG	50.7	59.7
OAB	42.6	48.9
TLD	51.8	67.3
CXT	**57.3**	65.7
BSBT	40.6	47.0
Our framework	51.9	**68.3**

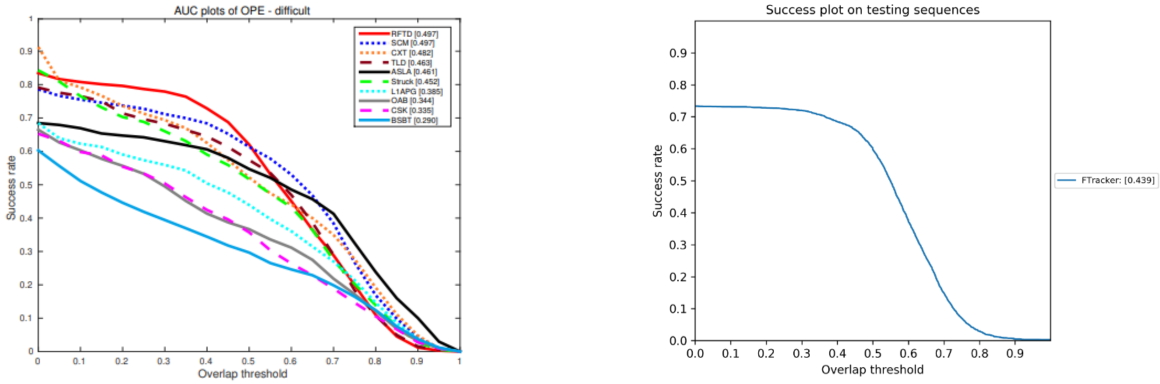
TABLE. V. TOP TRACKER COMPARISON ON OTB DATASET FACE SUBSET (DIFFICULT SET). EVALUATED RESULTS ARE FROM RFTD PAPER [46]

Face Tracker	Success Plot AUC	Success plot Threshold (0.5)
RFTD	**49.7**	**62.0**
Struck	45.2	51.7
SCM	**49.7**	**61.3**
ASLA	46.1	54.7
CSK	33.5	52.2
L1APG	38.5	43.9
OAB	34.4	36.6
TLD	46.3	57.4
CXT	**48.2**	52.2
BSBT	29.0	29.7
Proposed framework	43.9	**59.7**



(a) (b)

Fig. 11. Our Normalised Precision Plot on OTB Dataset Face Subsets (a) Normal Set (b) Difficult Set.



(a) (b)

Fig. 12. Success Plots of Trackers on OTB Dataset Face Subset (difficult set): (a) Results from RFTD Paper[46] (b) Results on our Method.

As first sight from the metric Table IV and Table V, the proposed framework has average AUC while the slope of the proposed framework is also shallow as predicted. The main reason here is because when the predicted box is drifted from the face, the algorithm terminates the tracklet instantly; therefore, with high normalised error, our tracker performs the same as with low normalised error while other trackers yield noticeably different results with different normalised errors. The initial modest success rate leads to a modest average value. The success rate at threshold 0.5 is still good, ranking third in that section in both subsets.

V. Conclusions

In this work, the authors proposed a method for face tracking problem in semi-online manner - the online process with some minor delay. The comparing experiments are conducted on two datasets: MobiFace dataset and OTB dataset with many state-of-the-arts works in the field. The results show that our method can produce robust accuracy while keeping a good speed. With that, the effectiveness of adding the tracklet-tracklet association stage after detection stage in semi-online manner is proven. The manipulation of appearance affinity and motion affinity have brought us the accuracy of the framework, while the workload division and information sharing of the two main stages make our process lighter and achieve better speed. With the improvements, all the disadvantages pointed out in Section I are solved.

The demonstrated framework has many advantages that can be applied to the production environment. First, the process as a whole was cut off to achieve the speed which is suitable for continuous streaming with a little delay. Second, the accuracy maintains at an acceptable value, which makes the proposed framework robust in many unconstraint environments. Finally, the framework can work without supervision, and is a high-performance multi-face tracking system.

There are many ways to develop from this work. First, because the framework consists of many components, researchers can try other combinations of related techniques (detector, tracker, feature extractor) to achieve better results. Second, the concept of semi-online tracking (use some delay for better results) can be applied to current work on face tracking.

Acknowledgment

This research is funded by Viet Nam National University Ho Chi Minh City (VNUHCM) under grant no. B2018-18-01.

Thank to Axon company for the valuable support on cooperation.

References

[1] J. Berclaz, F. Fleuret, E. Türetken, and P. Fua, “Multiple Object Tracking Using K-Shortest Paths Optimization,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, pp. 1806–1819, 2011.

[2] A. Milan, S. Roth, and K. Schindler, “Continuous Energy Minimization for Multitarget Tracking,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 1, pp. 58–72, Jan. 2014.

[3] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg, “Multiple Hypothesis Tracking Revisited,” 2015 IEEE Int. Conf. Comput. Vis. ICCV, pp. 4696–4704, 2015.

[4] S. Tang, B. Andres, M. Andriluka, and B. Schiele, “Multi-Person Tracking by Multicut and Deep Matching,” ArXiv E-Prints, p. arXiv:1608.05404, Aug. 2016.

[5] L. Leal-Taixé, C. Canton Ferrer, and K. Schindler, “Learning by tracking: Siamese CNN for robust target association,” ArXiv E-Prints, p. arXiv:1604.07866, Apr. 2016.

[6] C. Cruz, L. Sucar, and E. Morales, “Real-Time face recognition for human-robot interaction,” in Proceedings of the 8th IEEE International Conference on Automatic Face and Gesture Recognition, 2008, pp. 1–6.

[7] A. V. Segal and I. D. Reid, “Latent Data Association: Bayesian Model Selection for Multi-target Tracking,” 2013 IEEE Int. Conf. Comput. Vis., pp. 2904–2911, 2013.

[8] A. Sadeghian, A. Alahi, and S. Savarese, “Tracking The Untrackable: Learning To Track Multiple Cues with Long-Term Dependencies,” ArXiv E-Prints, p. arXiv:1701.01909, Jan. 2017.

[9] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu, “Online Multi-Object Tracking Using CNN-based Single Object Tracker with Spatial-Temporal Attention Mechanism,” ArXiv E-Prints, p. arXiv:1708.02843, Aug. 2017.

[10] L. Chen, H. Ai, Z. Zhuang, and C. Shang, “Real-time Multiple People Tracking with Deeply Learned Candidate Selection and Person Re-Identification,” ArXiv E-Prints, p. arXiv:1809.04427, Sep. 2018.

[11] C. Kim, F. Li, and J. M. Rehg, “Multi-object Tracking with Neural Gating Using Bilinear LSTM,” in ECCV, 2018.

[12] M. Thoreau and N. Kottege, “Improving Online Multiple Object tracking with Deep Metric Learning,” ArXiv E-Prints, p. arXiv:1806.07592, Jun. 2018.

[13] Y. Yoon, A. Boragule, Y. Song, K. Yoon, and M. Jeon, “Online Multi-Object Tracking with Historical Appearance Matching and Scene Adaptive Detection Filtering,” ArXiv E-Prints, p. arXiv:1805.10916, May 2018.

[14] N. Narayan, N. Sankaran, S. Setlur, and V. Govindaraju, “Re-identification for Online Person Tracking by Modeling Space-Time Continuum,” 2018 IEEECVF Conf. Comput. Vis. Pattern Recognit. Workshop CVPRW, pp. 1519–151909, 2018.

[15] S. Zhang et al., “Improved Selective Refinement Network for Face Detection,” ArXiv E-Prints, p. arXiv:1901.06651, Jan. 2019.

[16] R. Kalman, “A New Approach to Linear Filtering and Prediction Problems,” Trans. ASME - J. Basic Eng., vol. 82, pp. 35–45, 1960.

[17] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “High-Speed Tracking with Kernelized Correlation Filters,” ArXiv E-Prints, p. arXiv:1404.7584, Apr. 2014.

[18] D. Held, S. Thrun, and S. Savarese, “Learning to Track at 100 FPS with Deep Regression Networks,” in Unknown, 2016, vol. 9905, pp. 749–765.

[19] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, “Fully-Convolutional Siamese Networks for Object Tracking,” ArXiv E-Prints, p. arXiv:1606.09549, Jun. 2016.

[20] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823.

[21] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, “Camera Style Adaptation for Person Re-identification,” ArXiv E-Prints, p. arXiv:1711.10295, Nov. 2017.

[22] J. Zhuo, Z. Chen, J. Lai, and G. Wang, “Occluded Person Re-identification,” ArXiv E-Prints, p. arXiv:1804.02792, Apr. 2018.

[23] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, “Part-Aligned Bilinear Representations for Person Re-identification,” ArXiv E-Prints, p. arXiv:1804.07094, Apr. 2018.

[24] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, “Human Semantic Parsing for Person Re-identification,” ArXiv E-Prints, p. arXiv:1804.00216, Mar. 2018.

[25] J. Almazán, B. Gajic, N. Murray, and D. Larlus, “Re-ID done right: towards good practices for person re-identification.,” CoRR, vol. abs/1801.05339, 2018.

[26] H. Wang et al., “CosFace: Large Margin Cosine Loss for Deep Face Recognition,” ArXiv E-Prints, p. arXiv:1801.09414, Jan. 2018.

[27] M. Keuper, E. Levinkov, N. Bonneel, G. Lavoué, T. Brox, and B. Andres, “Efficient Decomposition of Image and Mesh Graphs by Lifted Multicuts,” ArXiv E-Prints, p. arXiv:1505.06973, May 2015.

[28] V. Traag, L. Waltman, and N. J. van Eck, “From Louvain to Leiden: guaranteeing well-connected communities,” ArXiv E-Prints, p. arXiv:1810.08473, Oct. 2018.

[29] H. W. Kuhn, “The Hungarian Method for the Assignment Problem,” in 50 Years of Integer Programming, 2010.

[30] S. Zhang et al., “Tracking Persons-of-Interest via Adaptive Discriminative Features,” in Computer Vision – ECCV 2016, Cham, 2016, pp. 415–433.

[31] S. Jin, H. Su, C. Stauffer, and E. Learned-Miller, “End-to-End Face Detection and Cast Grouping in Movies Using Erdös-Rényi Clustering,” in arXiv e-prints, 2017, pp. 5286–5295.

[32] C. Lin and Y. Hung, “A Prior-Less Method for Multi-face Tracking in Unconstrained Videos,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 538–547.

[33] F. Comaschi, S. Stuijk, T. Basten, and H. Corporaal, “Online multi-face detection and tracking using detector confidence and structured SVMs,” in 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2015, pp. 1–6.

[34] P. Viola and M. J. Jones, “Robust Real-Time Face Detection,” Int. J. Comput. Vis., vol. 57, no. 2, pp. 137–154, May 2004.

[35] M. Naiel, M. O. Ahmad, M. N. s Swamy, J. Lim, and M.-H. Yang, “Online Multi-Object Tracking via Robust Collaborative Model and Sample Selection,” Comput. Vis. Image Underst., vol. 154, 2016.

[36] X. Lan, Z. Xiong, W. Zhang, S. Li, H. Chang, and W. Zeng, “A super-fast online face tracking system for video surveillance,” in 2016 IEEE International Symposium on Circuits and Systems (ISCAS), 2016, pp. 1998–2001.

[37] A. Ranftl, F. Alonso-Fernandez, S. Karlsson, and J. Bigun, “Real-time AdaBoost cascade face tracker based on likelihood map and optical flow,” IET Biom., vol. 6, no. 6, pp. 468–477, 2017.

[38] J. Chen, R. Ranjan, A. Kumar, C. Chen, V. M. Patel, and R. Chellappa, “An End-to-End System for Unconstrained Face Verification with Deep Convolutional Neural Networks,” in 2015 IEEE International Conference on Computer Vision Workshop (ICCVW),2015,pp.360–368.

[39] M. Hayat, S. H. Khan, N. Werghi, and R. Goecke, “Joint Registration and Representation Learning for Unconstrained Face Identification,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1551–1560.

[40] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, and A. Zisserman, “Template Adaptation for Face Verification and Identification,” in 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), 2017, pp. 1–8.

[41] R. Ranjan et al., “A Fast and Accurate System for Face Detection, Identification, and Verification,” IEEE Trans. Biom. Behav. Identity Sci., vol. 1, pp. 82–96, 2018.

[42] Y. Wang, J. Shen, S. Petridis, and M. Pantic, “A real-time and unsupervised face Re-Identification system for Human-Robot Interaction,” Pattern Recognit. Lett., 2018.

[43] Jianbo Shi and Tomasi, “Good features to track,” in 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1994, pp. 593–600.

[44] Y. Lin, S. Cheng, J. Shen, and M. Pantic, “MobiFace: A Novel Dataset for Mobile Face Tracking in the Wild,” ArXiv E-Prints, p. arXiv:1805.09749, May 2018.

[45] Y. Wu, J. Lim, and M.-H. Yang, “Object Tracking Benchmark,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, pp. 1–1, 2015.

[46] F. Comaschi, S. Stuijk, T. Basten, and H. Corporaal, “Robust online face tracking-by-detection,” in 2016 IEEE International Conference on Multimedia and Expo (ICME), 2016, pp. 1–6.

Mobile Sensor Node Deployment Strategy by using Graph Structure based on Estimation of Communication Connectivity and Movement Path

Koji Kawabata[1], Tsuyoshi Suzuki[2]
Department of Information and Communication
Engineering, Tokyo Denki University
Tokyo, Japan

***Abstract*—We propose a multiple-mobile sensor node (MSN) deployment strategy that considers wireless communication quality and operation time of underground wireless sensor networks. After an underground disaster, it is difficult to perform a rescue operation because the internal situation cannot be confirmed. Hence, gathering information using a teleoperated robot has been widely discussed. However, wireless communication is unstable and the corresponding wireless infrastructure to operate the teleoperated robot is unavailable underground. Therefore, we studied the disaster information-gathering support system using wireless sensor networks and a rescue robot. In this study, the movement path information of the teleoperated robot is fed to MSNs in a graph structure. MSNs are deployed in the underground environment by adding an evaluation of communication quality and operation status to a given graph structure. The simulation was evaluated in an assumed underground environment. The results confirmed that the wireless communication quality between each MSN was maintained and energy consumption was balanced during the deployment.**

Keywords—Wireless sensor networks; deployment strategy; communication connectivity

I. Introduction

Gathering information in disaster areas is very important for assessing the situation, avoiding secondary disasters, and reducing disasters [1–5]. In general, bird's-eye image information gathered by unmanned air vehicles (UAVs) and artificial satellites is useful for understanding post-disaster situations. However, in an underground space in the city where such UAVs cannot gather information, it is difficult to ascertain the extent of the damage, which is important for avoiding secondary disasters. In addition, rescue teams cannot organize a suitable rescue plan for underground spaces owing to the lack of sufficient information. In these situations, the rescue team go underground to gather disaster information. The information is then shared within the on ground and underground teams for efficient and cooperative rescue work. However, when the communication infrastructure is broken due to damage, rescue teams cannot cooperate closely because of the disconnection. Therefore, the rescue team must work underground without the complete knowledge of the situation and face the added risk of secondary disasters.

From studies based on past accident analysis, researchers have recently focused on a disaster information-gathering method using a wireless sensor network (WSN) and a rescue robot in closed areas. The WSN consists of spatially distributed sensor nodes (SNs) to cooperatively monitor environmental conditions such as temperature, sound, vibration, pressure, motion, etc. The WSN is then enabled to provide wireless communication without the existing infrastructure. In a closed area, it is constructed using a rescue robot. Therefore, an information-gathering method by constructing the communication infrastructure in a disaster area using a WSN has been discussed [6,7]. However, the scope of application is limited to outside the disaster area, and studies assuming a closed space have not been reported.

Information gathering by rescue robots and disaster rescue support systems is effective. The use of robot technology can reduce the activity burden on rescue workers. Rescue robots are often remotely controlled by considering the impact on the disaster area and work safety. When the robot is operated remotely, it is possible to support rescue operations in spaces where people cannot easily enter, such as closed spaces, narrow spaces, and underwater. The connectivity and stability of communication are very important during remotely controlled operations. Owing to the closed environment, wireless communication is often unstable underground, as compared to the outdoors. Hence, degradation of wireless communication due to disturbances such as fading and shadowing is more likely to occur underground than outdoors.

Therefore, we studied the information-gathering system using teleoperated robots and WSNs, as observed in Fig. 1 [8–10]. In this system, a WSN is constructed using a mobile sensor node (MSN). This system responds quickly to network disconnection. This system reports that the end-to-end throughput was maintained and an effective WSN was constructed. However, the autonomous deployment of each MSN is unavailable in this system. In environments where wireless communication is unstable, such as the underground, the MSN must be placed in a position that provides a stable wireless relay to maintain the communication connection for the teleoperated robot. In addition, the MSN has limited energy because the power supply is a battery. Hence, it is important for the WSN that there is no network disconnection when the MSN stops operating.

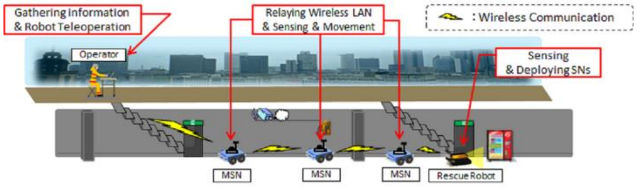


Fig. 1. Disaster Information-Collecting Support System using a Remote-Control Robot and WSNs.

Certain studies reported on relocating the MSN to be positioned where communication quality was evaluated in the field of the WSN [11,12]. Furthermore, in the field of multi-agent systems, cooperative control of multiple robots based on communication quality has been reported [13,14]. These studies utilized the received signal strength indicator (RSSI) for communication quality. Saitou [13] was reported that the deployment control which made always uniform RSSI between each robot was performed. Zhong [14] was presented that the distributed control which establish and maintain RSSI between the robot was performed. By evaluating RSSI, it is demonstrated that the MSN can be autonomously deployed at the position where wireless communication between MSNs is stabilized in WSNs.

In general, WSNs are required to operate for a long time and hence, energy efficient relocation methods have been reported [15,16]. The relocation of the SNs at the appropriate movement distances reduced the network energy consumption and maximized the coverage level. Conversely, studies have proposed that each MSN is placed at the position where the energy consumption between each MSN is balanced [17,18]. By applying energy costs, such as energy consumption and the standard deviation of energy consumption between MSN to coverage to MSN deployment model, the overall energy of WSNs is balanced. These studies conclude that long-term operation is possible by reducing and balancing the energy consumption of each SN.

Therefore, the proposed MSN deployment strategy considers an underground space, which is positioned where the communication quality between MSNs and the operating status of the MSN are evaluated. We then evaluated the communication quality between MSNs and energy balance of each MSN using a simulation.

II. Deployment Strategy of MSN

Our proposed deployment strategy requires a graph structure construction and autonomous operation of multiple MSNs. In a disaster environment where the situation is unknown, it is difficult to remotely control a robot and multiple MSNs simultaneously. Therefore, it is desirable to position MSNs where communication quality in the environment is maintained while moving autonomously. However, a fully autonomous operation of the robot may malfunction in the disaster-area environment.

Hence, the robot is operated remotely such that when a movement target is given by the operator, the movement path information is shared by the WSN. The MSN is autonomously operated by following the robot based on its movement path information. In addition, each MSN evaluates communication quality and its own operation status while moving on the path and is placed on the path based on the evaluation.

A. Precondition

In the information-gathering system, the network is constructed in the order of the operator, MSNs, and robot in a straight line. Furthermore, a change in the network topology during the construction of the WSN may disconnect the network when the communication path is changed. Particularly, wireless communication in an underground space is unstable, and it is likely for the network to disconnect. Therefore, the network topology of this system does not change.

Information sharing with neighboring terminals is necessary for autonomous operation of MSNs. The information shared with the neighboring terminal is (1) the three-dimensional absolute coordinate information, (2) a remaining energy, and (3) path information of robot movement. The MSN is operated to follow the robot’s path. Each MSN acquires the coordinate information of the neighboring terminals to accurately follow the robot. Further, the remaining energy is used to evaluate the operating status. Because the WSN is constructed in a straight line, each MSN shares information with two terminals. Furthermore, the information shared is regularly updated. The update frequency of these information differs. The information of (1) and (2) is regularly updated 1 seconds because MSN follow the robot’s path accurately. The information of (3) is regularly updated 10 seconds because the different received information is assumed.

B. Construction Guidelines for a Graph Structure

The walls and ceilings are less likely to collapse during a disaster because the underground space has a strong structure and an excellent seismic resistance [19]. Therefore, it is unlikely that the environmental structure will change significantly. However, it is difficult to use the map when obstacles occur due to the disaster situation. To place multiple MSNs depending on the environmental situation, it is necessary to analyze the environment. Because SLAM (Simultaneous Localization and Mapping) is often used to analyze the environment by a mobile robot, the teleoperated robot also uses SLAM to collect information while analyzing the environment, which builds the detailed environment with a metric map. In contrast, when MSN requires the robot to follow its path, it is easier to control the robot by specifying a characteristic position on the movement path. An example of a characteristic position is the inflection point of a passage.

Therefore, a topological map is used to represent the environmental information given to each MSN. The topological map represents the environment in a graph structure as shown in Fig. 2. A topological map enables adaptive movement control to the environment while saving computational resources of the MSN. The topological map can plan a path by providing coordinate information to a node [20,21]. The nodes are set to a position where the communication quality is likely to fluctuate and the MSN is relatively difficult to move. For example, the node is set to a coordinate that starts and ends at a stair, such as a corner or a crossroad. The edge is set on the direct movement path between the robot and MSN.

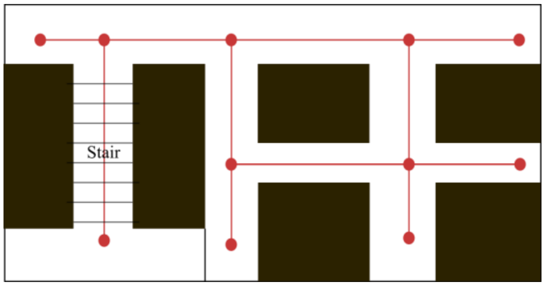


Fig. 2. Processing Flow of the Deployment Algorithm.

The topological map is constructed using sensors (sonar and LRF (Laser Range Finder)) mounted on the teleoperated robot. When a characteristic position is observed by the robot sensors, the node is set to absolute coordinates of the robot at that position. By setting the cost of the edge to the communication quality, the MSN can move along the path where communication is stable. Because each MSN shares the same map, the graph structure is sent to each MSN and the base station.

C. Deployment Algorithm for MSN

1) Algorithm overview: Fig. 3 shows the MSN deployment algorithm. Each MSN periodically measures the communication quality between neighboring MSNs to evaluate the communication quality. When detecting a decrease in the quality, the detected MSN evaluates the remaining energy between neighboring MSNs. Based on the evaluation value, the MSN with the large remaining energy, i.e., the MSN with a sufficient operating time moves preferentially. The MSN to be moved calculates the moving direction from the shortest path information and moves. It stops when the communication quality recovers. With this algorithm, each MSN is operated autonomously while maintaining the communication quality between the MSNs and balancing the energy of each MSN.

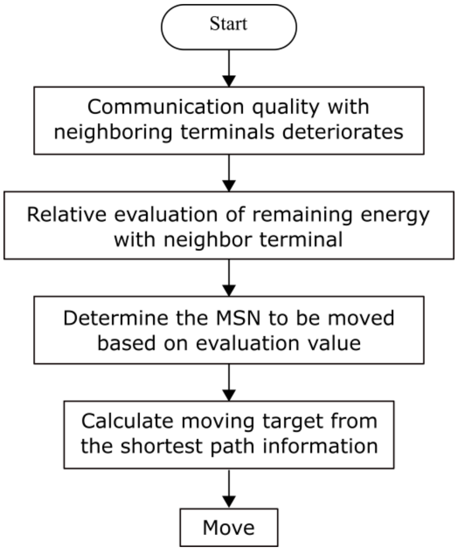

Start
Communication quality with neighboring terminals deteriorates
Relative evaluation of remaining energy with neighbor terminal
Determine the MSN to be moved based on evaluation value
Calculate moving target from the shortest path information
Move

Fig. 3. Topological Map Assuming an underground Space. The Node is Set to the Coordinates at the Start and end of a Stair, such as a Corner or a Crossroad.

2) Path planning: The path planning based on the topological map determines the candidate of the deployment position by calculating the shortest path between each MSN. To perform remote control between the operator and the robot, it is necessary to place each MSN at a position that maintains communication connectivity on the movement path based on the map. Therefore, map-based path planning is necessary. It determines the candidate of the deployment position by calculating the shortest path. The algorithm used in path planning uses Dijkstra's algorithm [22]. Path planning is performed at regular intervals to always obtain the latest movement path information. The update frequency is 10 seconds, which is the same as the interval for acquiring the movement path information from neighboring MSN. Each MSN selects the nodes up to neighboring MSN by path planning and stores the coordinates of the node in the data memory. The MSN moves by referring to the coordinates from its own data memory.

3) Evaluation of communication quality: The communication index uses RSSI. Each MSN sets a threshold to detect a decrease in the RSSI. The MSN determines whether to relocate based on the threshold. When an MSN detects that at least one RSSI has fallen below the threshold, it shifts to the process of evaluating the remaining energy. The MSN evaluates the remaining energy for the neighboring MSN connected to the communication path with the worst RSSI. If the RSSI of the multiple communication paths fall below the threshold, the MSN prioritizes recovering the minimum RSSI. The RSSI is maintained by relocating the MSNs to a position where the RSSI does not always decrease.

4) Evaluation of remaining energy: Each MSN evaluates the remaining energy with the neighboring MSN. The MSN with the higher remaining energy moves preferentially such that the energy consumption of each MSN is uniform. *The* evaluation of the remaining energy is relative and is performed by the MSN that detects a decrease in RSSI. The relative evaluation of the remaining energy is calculated using

$$\varepsilon = \frac{|E(i) - E(j)|}{\max(E(i),\ E(j))}\ (0 \leq\ \varepsilon \leq 1) \quad (1)$$

$E(i)$ is the remaining energy of the MSN performing the evaluation process and $E(j)$ is the remaining energy of the neighboring MSN that is to be evaluated. If the value of ε is closer to 1, the difference between the remaining energy of $E(i)$ and $E(j)$ is large. Further, both the MSNs that are to be evaluated are determined the moving MSN based on the evaluation value. The MSN is determined according to the following conditions:

- When $\varepsilon \leq \varepsilon_{th}$, the evaluated MSN moves.
- When $\varepsilon > \varepsilon_{th}$, the MSN of $\max(E(i),\ E(j))$ moves. The other MSN does not move.

ε_{th} is the threshold of the evaluation value of the remaining energy. After the MSN to be moved is determined by this condition, the MSN starts moving.

5) Movement control of MSN based on the topological map: The MSN moves based on the coordinates of the nodes obtained in the shortest path to recover the RSSI. Fig. 4 demonstrates the MSN relocation due to the decrease in RSSI. $RSSI_{th}$ is the RSSI threshold between neighboring terminals. As previously mentioned, when $RSSI_{th}$ was lower than RSSI, the MSN to move was determined by the relative evaluation of the remaining energy. Furthermore, the moving MSN refers to the coordinates from the data memory used to recover the RSSI. These coordinates are defined as sub-targets. The MSN moves by calculating the moving direction from the relative coordinates with the sub-target. When the MSN reaches the sub-target, the next coordinate is referred to from the data memory and again set as a new sub-target. It then calculates the moving direction and moves. If the RSSI exceeds the threshold before the MSN reaches the sub-target, the MSN stops at that position.

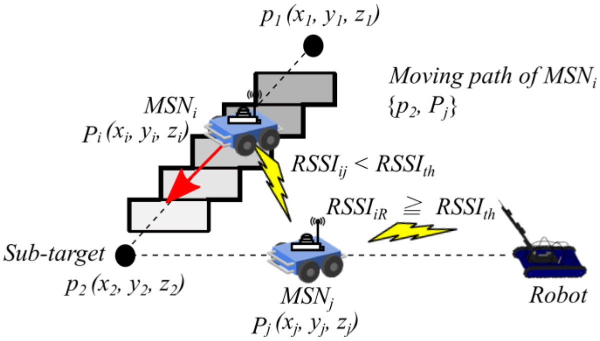


Fig. 4. MSN_i Recovers RSSI by Running Down Stairs based on the Shortest Path.

III. Simulations

The simulation of the proposed method was verified in a three-dimensional coordinate-system environment without obstacles. The simulator used the open-source network simulator ns-3 [23], which has a model of wireless communication and radio wave propagation of mobile communication bodies. In the simulation, the base station is defined as Sink and the teleoperated robot as Leader.

The simulation conditions are shown in Table I. The network topology is constructed linearly with Sink－MSN1－MSN2 － MSN3 － Leader and is not changed during the simulation. Network settings are set according to the wireless communication protocol of the information collection system which is being discussed. Radio wave propagation is based on a noise model. Adding a noise model to the simulation helps in verifying the method in a realistic environment. The method's reliability can also be improved. Fig. 5 depicts the graph structure used in the experiment. The nodes and edges in this experiment were verified using a pre-built graph structure. The nodes are assigned IDs and have coordinates. The cost of an edge is the linear distance between the nodes. The nodes were set at these positions in accordance with the universal underground structure in Japan. The simulation, we demonstrated the effectiveness of the proposed method by performing two verification experiments.

TABLE. I. Simulation Setting

Parameter	***Unit***	***Value***
Number of MSN		4
Initial position (Sink MSN1, MSN2, MSN3, Leader)	(m)	(-0.8, -0.8, 0) (-0.6, -0.6, 0) (-0.4, -0.4, 0) (-0.2, -0.2, 0) (0, 0, 0)
Link-layer protocol		IEEE802.11g
Network/Transport protocol		UDP/IP
Radio-wave propagation loss model		Free space
Noise model		Nakagami m distribution
Node velocity	(m/s)	0.2
Tx power	(dBm)	16
RSSI threshold	(dBm)	-50

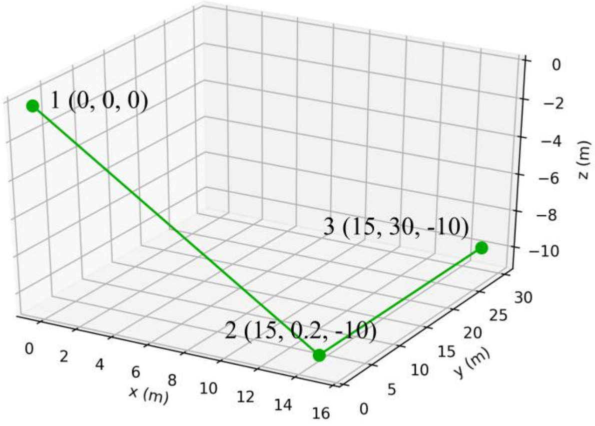


Fig. 5. Graph Structure using Experiment. These Nodes are Assigned IDs and have Coordinates.

A. Evaluation of RSSI by Deployment Algorithm

This experiment evaluated RSSI between each MSN over time. The Leader moves in the order of nodes with IDs 1, 2, and 3. Each MSN moves through the node of the shortest path calculated by path planning.

The results are shown in Fig. 6 and 7. Fig. 6 represents the RSSI between the deployed MSNs when threshold of the RSSI is -50 dBm, and. Fig. 7 shows the movement trajectory of each MSN. In Fig. 6, Sink－MSN1, MSN1－MSN2, MSN2－MSN3, MSN3－Leader are the observed RSSI between each MSN. Sink－MSN1 (10 s MA), MSN1－MSN2 (10 s MA), MSN2－MSN3 (10 s MA), MSN3－Leader (10 s MA) are the calculated RSSI by moving average every 10 seconds. Each RSSI was maintained at approximately -35 dBm for Sink－MSN1, approximately -50 dBm for MSN1－MSN2 and MSN2 －MSN3, and approximately -62 dBm for MSN3－Leader. Each RSSI was observed to converge in 300 s. In Fig. 7, each MSN was deployed according to the graph structure. Each MSN moved within MSN1 at (1.165, 0.016, -0.777), MSN2 at (8.486, 0.113, -5.658), and MSN3 at (15, 1.4, -10).

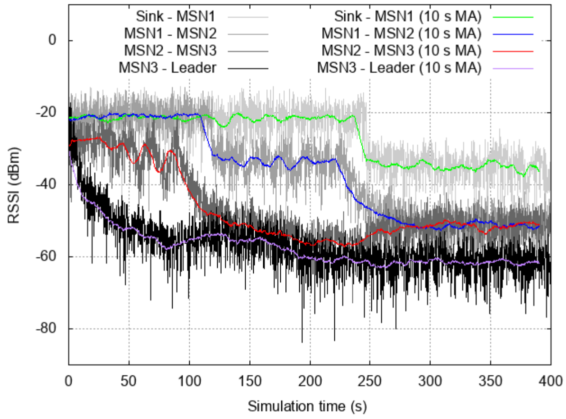


Fig. 6. RSSI During Deployment over Time.

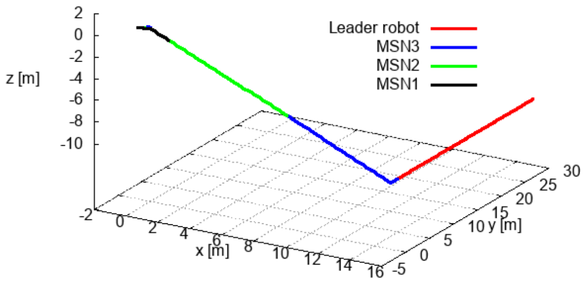


Fig. 7. Movement Trajectory of each MSN.

B. Comparison of Moving Energy Consumption

This experiment was compared with the deployment algorithm used in this study when the remaining energy was not evaluated. We evaluated the total mobile energy consumption when each MSN was deployed. Based on the results, we performed relative evaluation of the remaining energy by (1) and evaluated the balanced energy between MSN. Table II lists the setting conditions for calculating energy consumption. The total energy consumption of each MSN was calculated by the sum of the energies of wireless communication, movement, and sensor. The movement of the Leader is the same as in the previous experiment.

The results are listed in Tables III and IV. Table III demonstrates the change in each MSN energy consumption when the process of evaluating the remaining energy is applied to the deployment algorithm and when it is not applied. Table IV shows the results of relative evaluation of the remaining energy of each MSN from Table III using (1). In Table III, the difference in energy consumption of each MSN increased by +1245 J for MSN1, but decreased by -1145 J for MSN2 and -1195.2 J for MSN3. In Table IV, the method that evaluates RSSI and remaining energy has a smaller relative evaluation value. If the relative evaluation value is closer to 0, the difference between the remaining energy of both MSNs is more balanced.

Here, we examined whether a difference appears in the convergence position by these comparison methods. The examination results are shown in Fig. 8. The red dot represents MSN1 at (1.165, 0.016, -0.777), MSN2 at (7.155, 0.095, -4.770), and MSN3 at (15, 0.6, -10). The blue dot represents MSN1 at (1.331, 0.017, -0.888), MSN2 at (7.654, 0.102, -5.103), and MSN3 at (15, 0.2, -10).

TABLE. II. ENERGY CONSUMPTION SETTING

Parameter	***Unit***	***Value***
Initial energy	(J)	100,000
Controller supply voltage	(V)	5.0
Motor supply voltage	(V)	15.0
Idle current	(A)	0.60
Tx current	(A)	0.85
Rx current	(A)	0.65
Motor current	(A)	3.22
Sensor energy consumption	(J)	2.5
ε_{th}		0.04

TABLE. III. ENERGY CONSUMPTION OF MSNS

MSN No.	**Moving energy consumption (J)**	
	Proposed method	***Proposed method (the remaining energy is not evaluated)***
1	5989.58	4744.58
2	9702.49	10847.89
3	15321.59	16516.79

TABLE. IV. ENERGY CONSUMPTION OF MSNS

	Relative evaluation value of remaining energy	
	Proposed method	***Proposed method (the remaining energy is not evaluated)***
MSN1 - MSN2	0.040	0.065
MSN2 - MSN3	0.063	0.064

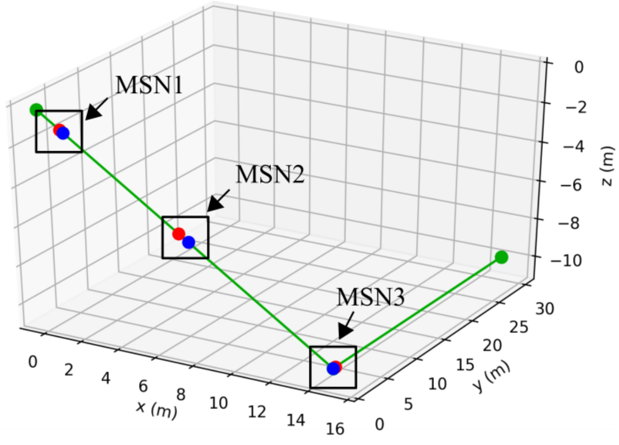


Fig. 8. Comparison of the Convergence Position of each MSN by Evaluation Method. The Red Dots are the Convergence Coordinates of each MSN by the Proposed Method, the Blue Dots are the Convergence Coordinates of each MSN by the Proposed Method (the Remaining Energy is not Evaluated).

IV. DISCUSSION

The simulations demonstrated that the proposed method could place multiple MSNs in a location where RSSI maintenance and energy consumption were balanced, including an environment with elevation differences. Thus, this disaster information-gathering support system can simultaneously achieve high reliability of communication connectivity and long-term operation. The RSSI observation resulted in instantaneous fluctuations due to noise. The MSN moves when the RSSI value at the time is less than the threshold. However, the MSN is at the position where the RSSI can be maintained without movement. As this movement is unnecessary, considering it might be inefficient in terms of energy and movement efficiency. Further, if the MSN is placed by considering the communication quality, noise countermeasures are necessary.

The moving energy consumption of the MSN1 increased because the operation load was concentrated on the MSN1 by the proposed method. The remaining energy of each MSN was balanced as a result. Moreover, since the comparison of the convergence positions of each MSN hardly changed, the reliability of energy consumption experiment was presented. Equation (1) can be used when the network topology is linear. In other words, if the network topology changes during the construction of the WSN, it is necessary to change (1) accordingly. Thus, the method of deploying multiple MSNs can solve two problems in this information-gathering system that maintains communication connectivity and operates for a long time by evaluating communication quality and energy.

V. Conclusion

This paper proposed a method of operating a robot and multiple MSNs in a disaster area information collection support system while considering communication quality and operating status. While collecting environment information using a robot, the information is represented by a topological map and output as the MSNs' moving route information. In addition, MSN was operated by evaluating RSSI and the remaining energy. The experimental results confirmed that deployment is possible in an environment with a difference in elevation by utilizing movement path information. Therefore, this strategy constructs stable WSNs that can operate for a long time using a rescue robot. The strategy is effective for gathering disaster-area information in actual disaster scenarios. The proposed strategy will be applied for WSN deployment in an actual underground space in the future.

Acknowledgment

We would like to thank Editage (www.editage.com) for English language editing.

References

[1] H. Kawakata, Y. Kawata, H. Hayashi, T. Tanaka, K. C. Topping, K. Yamori, P. Yoshitomi, and T. Kugai, "Building an integrated database management system of information on disaster hazard, risk, and recovery process," Annuals of Disas. Prev. Res. Inst., Kyoto Univ., No. 47 C, 2004.

[2] Y. Kawata, "Disaster mitigation due to next Nankai earthquake tsunamis occurring in around 2035," Proceedings of International Tsunami Symposium 2001, session 1, pp. 315–329, 2001.

[3] Y. Kawata, "The great Hanshin-Awaji earthquake disaster: damage, social response, and recovery," Journal of Natural Disaster Science, vol. 17, no. 2, pp. 1–12, 1995.

[4] M. Erdelj, E. Natalizio, K. R. Chowdhury, and I. F. Akyildiz, "Help from the Sky: Leveraging UAVs for Disaster Management," IEEE Pervasive Computing, vol. 16, no. 1, pp. 24–32, 2017.

[5] W. Shan, J. Feng, J. Chang, F. Yang, and Z. Li, "Collecting earthquake disaster area information using smart phone," 2012 International Conference on System Science and Engineering (ICSSE), Dalian, Liaoning, pp. 310–314, 2012.

[6] T. Gurkan, G. V. Cagri, and G. Kayhan, "An autonomous wireless sensor network deployment system using mobile robots for human existence detection in case of disasters," Ad Hoc Networks, vol. 13, pp. 54-68, 2014.

[7] A. Mangla, A. K. Bindal, and D. Parasad, "Disaster Management in wireless sensor networks: A survey report," International Journal of Computing and Corporate Research, vol. 6, 2016.

[8] K. Sawai, T. Suzuki, H. Kono, Y. Hada, and K. Kawabata, "Development of a SN with impact-resistance capability for gathering disaster area information," 2008 International Symposium on Nonlinear Theory and its Applications (NOLTA2008), pp. 17–20, 2008.

[9] T. Suzuki, R. Sugizaki, K. Kawabata, Y. Hada, and Y. Tobe, "Autonomous deployment and restoration of sensor network using mobile robots," International Journal of Advanced Robotic System, vol. 7, no. 2, pp. 105–114, 2010.

[10] K. Sawai, S. Tanabe, H. Kono, Y. Koike, R. Kunimoto, and T. Suzuki, "Construction strategy of wireless sensor networks with throughput stability by using mobile robot," International Journal of Advanced Computer Science and Applications, The Science and Information organization, vol. 5, no. 2, pp. 14–20, 2014.

[11] M. Rajesh, A. George, and T. S. B. Sudarshan, "Energy efficient deployment of wireless sensor network by multiple mobile robots," 2015 International Conference on Computing and Network Communications (CoCoNet), pp. 72–78, 2015.

[12] Z. Hao, N. Qu, X. Dang, and J. Hou, "RSS-based coverage deployment method under probability model in 3D-WSN," in IEEE Access, vol. 7, pp. 183091–183104, 2019.

[13] T. Saitou, M. Nukada, and Y. Uchimura, "Deployment control of mobile robots for wireless network relay based on received signal strength," 2013 IEEE Workshop on Advanced Robotics and its Social Impacts, pp. 237–242, 2013.

[14] X. Zhong and Y. Zhou, "Maintaining wireless communication coverage among multiple mobile robots using Fuzzy Neural Network," IEEE/ASME International Conference on Mechatronics and Embedded Systems and Applications (MESA), pp. 35-41, 2012.

[15] J.Wang, C. Ju, Y. Gao, A. K. Sangaiah, and G. J. Kim, "A PSO based energy efficient coverage control algorithm for wireless sensor networks," Comput. Mater. Contin, Vol. 56, No. 3, pp. 433–446, 2018.

[16] M. Bakshi, A. Ray, and D. De, "Maximizing lifetime and coverage for minimum energy wireless sensor network using corona based sensor deployment," CSI transactions on ICT, Vol. 5, No. 1, pp. 17–25, 2017.

[17] Y. Li, B. Zhang, and S. Chai, "An energy balanced-virtual force algorithm for Mobile-WSNs," 2015 IEEE International Conference on Mechatronics and Automation (ICMA), pp. 1779–1784, 2015.

[18] S. Halder and A. Ghosal, "Is sensor deployment using Gaussian distribution energy balanced?" 2014 IEEE 11th Consumer Communications and Networking Conference (CCNC), pp. 721–728, 2014.

[19] J. P. Godard, "Urban underground space and benefits of going underground," Proc. World Tunnel Congress 2004 and 30th ITA General Assembly, Singapore, pp. 1–9. May 22–27 ，2004.

[20] H. Cheng, H. Chen, and Y. Liu, "Topological indoor localization and navigation for autonomous mobile robot," IEEE Transactions on Automation Science and Engineering, vol. 12, no. 2, pp. 729–738, 2015.

[21] Y. Kato, K. Kamiyama, and K. Morioka, "Autonomous robot navigation system with learning based on deep Q-network and topological maps," 2017 IEEE/SICE International Symposium on System Integration (SII), pp. 1040–1046, 2017.

[22] E. W. Dijkstra, "A note on two problems in connexion with graphs," Numerische Mathematik，Vol. 1, No. 1, pp. 269-271, 1959.

[23] ns-3 Network Simulator, https://www.nsnam.org/ (Accessed 2020-2-18)

Classification of Malignant and Benign Lung Nodule and Prediction of Image Label Class using Multi-Deep Model

Muahammad Bilal Zia[1], Zhao Juan Juan*[2]
Ning Xiao[4], Jiawen Wang[5], Ammad Khan[6]
School of Information and Computer
Taiyuan University of Technology, Taiyuan, China

Xujuan Zhou[3]
School of Management and Enterprise
University of Southern Queensland
Toowoomba, Australia

***Abstract*—Lung cancer has been listed as one of the world's leading causes of death. Early diagnosis of lung nodules has great significance for the prevention of lung cancer. Despite major improvements in modern diagnosis and treatment, the five-year survival rate is only 18%. Before diagnosis, the classification of lung nodules is one important step, in particular, because automatic classification may help doctors with a valuable opinion. Although deep learning has shown improvement in the image classifications over traditional approaches, which focus on handcraft features, due to a large number of intra-class variational images and the inter-class similar images due to various imaging modalities, it remains challenging to classify lung nodule. In this paper, a multi-deep model (MD model) is proposed for lung nodule classification as well as to predict the image label class. This model is based on three phases that include multi-scale dilated convolutional blocks (MsDc), dual deep convolutional neural networks (DCNN A/B), and multi-task learning component (MTLc). Initially, the multi-scale features are derived through the MsDc process by using different dilated rates to enlarge the respective area. This technique is applied to a pair of images. Such images are accepted by dual DCNNs, and both models can learn mutually from each other in order to enhance the model accuracy. To further improve the performance of the proposed model, the output from both DCNNs split into two portions. The multi-task learning part is used to evaluate whether the input image pair is in the same group or not and also helps to classify them between benign and malignant. Furthermore, it can provide positive guidance if there is an error. Both the intra-class and inter-class (variation and similarity) of a dataset itself increase the efficiency of single DCNN. The effectiveness of mentioned technique is tested empirically by using the popular Lung Image Consortium Database (LIDC) dataset. The results show that the strategy is highly efficient in the form of sensitivity of 90.67%, specificity 90.80%, and accuracy of 90.73%.**

Keywords*—*Lung nodule classification; dilated blocks; dual DCNNs; multi-task learning; multi-deep model

I. Introduction

Lung cancer is the world's most prevalent and deadly type of cancer. The failure to diagnose the early stages of lung cancer is one cause of higher mortality induced by lung cancer because symptoms usually appear in the final stages [1]. In 2018, an estimated 154,000 deaths were recorded in the USA from lung cancer, which is 1/4 of all cancer death. In developed countries, there is a 16% chance for lung cancer patients of a five-year survival rate [2]. The detection of lung cancer in the initial phase is difficult due to dime-sized lesion growth. The small lesions can just detectable by computerized tomography (CT) scan, and it takes the amount of effort of radiologists to detect and label them as benign or malignant. Computer-aided diagnosis (CAD) systems help the radiologist to decrease the burden as a second option. Lung Nodule classification is also a challenging task in computer-aided diagnosis.

Nowadays, deep learning, which is rebranded from neural networks, is considered as one of the best solutions to solve many strenuous problems of computer vision and pattern recognition like medical diagnostics, natural language processing, etc.

Even though deep learning techniques perform better in the state-of-art in different medical images tasks, but still, there are many challenges to solve. For instance, facing issues with small medical dataset and biological variations. In [16,17], the pre-trained DCNN architectures have been used due to robust learning capability from large scale datasets like ImageNet to solve the generally small amount of data visual recognition problems.

The main issue of classifying lung nodules is inter-class ambiguity and intra-class variations [18], which pose complex challenges in different modalities in the differentiation of benign lesions from malignant ones. A clear example that shows the complication is shown in Fig. 1. In which there is a massive difference between (a) benign lesions, and (b) malignant lesions. Although both benign and malignant lesions are similar, respectively, in both color and shape. Even though several neural networks are productive sufficient to memorize all the training samples [19] due to useful ability of deep learning models, the uncertainty formed by intra-class variation and inter-class similarity can disturb the neural networks and make them fall into misperception.

*Corresponding Author

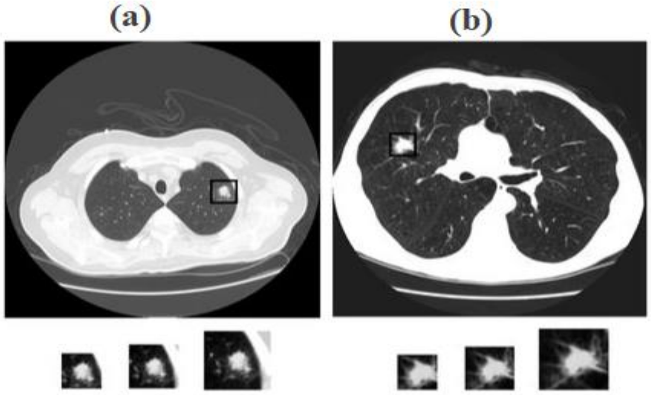


Fig 1. Example of (a) Benign nodule and (b), Malignant nodule.

To address the present challenges, in this paper, a multi-deep model is used for classification of a lung nodule in LIDC datasets, which consists of multi-scale dilated convolutional layer blocks, dual DCNNs, and multi-task learning component. The multi-scale method is applied on LIDC datasets, which helps to view and figure out the dataset at different scales. It worked as a speckle noise filter by eliminating the high-frequency constituent at every scale where no edges were spotted. Moreover, the dual pre-trained DCNNs extract the parameters from the ImageNet dataset [20], and these parameters were fine-tuned with our dataset. Both dual DCNNs simultaneously learn the representation of images from a pair of images, along with two similar images in various categories and two different images in the same category due to strong learning ability. The multi-task learning technique is applied on input pair of images where it will classify the multi-scale dilated Convolutional layers (MsDc) paired images which belong to the similar class or not and also helps to classify between malignant and benign. In addition, the multi deep model is easily trained “end-to-end” in classification supervision. During the test stage, each sample is given with the precision probabilities of a dual neural network together as the probability for a joint decision. The MD model is tested on LIDC datasets. Hence it is proved that the proposed model is state-of-the-art on lung nodule classification problem

II. Related Work

Nowadays deep learning is helping in many medical fields to improve image classification accuracy. From the last few years, a number of solutions are being published to solve the image classification problems [3, 4, 5, 6, and 7]. These solutions contain handcrafted feature extraction and classifier learning process. For just a classification task, it is very difficult to design such a handcrafted feature. So, deep learning models help for superior classification and to minimize the need for manual feature design. It also helps to enhance the performance of medical image detection, classification [8, 9], and segmentation [10, 11]. For instance, Jung H et al. [12] adopt a 3D-DCNN with dense connection and shortcut connection technique for lung nodule classification using LUNA 16 datasets to capture the 3D features. Xu Y et al. [13] use a deep convolutional neural network algorithm in their paper to minimize the manual annotation and produce superior feature portrayal for classification and segmentation of colorectal cancer using Cellular pathology. In machine learning, multi-task learning techniques are used in many applications like in drug detection [21], speech recognition [22], spatio-temporal event forecasting [28] and in natural language processing [23], etc. Li X et al. [14] proposed a multi-task learning framework that captures all the lung nodule assortment by taking out all the distinctive features using Convolutional Neural Network (CNN) from alternatingly stacked layers. For improvement in the final result, they train the CNN and form multi-tasking learning which shares information among nine different nodule features at the same time. Wu Z et al. [15] developed a multi-scale convolutional neural network (CNN) for removing lesion surface from CT scans which rely on 3D context fusion named M3DCF. An improved Multi-scale algorithm is used by Qingyuan [24] for image enhancement by using a canny operator to divide the image into edge region and non-edge region zone. Wang Y et al. [25] introduced a succinct and powerful multi-scale dilated convolutional method which used the dilated filters to integrate situational multi-scale data without reducing the receptive field efficiently. The logic behind this approach was to focus on the phenomenon that the dilated convolutional can efficiently extend the correct receptive area while retaining the useful contextual information. In the meantime, they also use residual methods to improve the learning process. Another dilated convolutional approach is used by Wei Y [26] in which they used several multiple dilated convolutional blocks with different dilated levels to create dense position maps of objects for weakly or semi-supervised manner for semantic segmentation networks. It has been also found that dilated convolutional [27], which by increasing the respective field size of kernels, offers a promising solution. In which they used different dilated levels and generated the localization maps at these dilated rates to enhance the discriminative ability.

III. Method

A. Multi-Deep Model

The proposed multi-deep (MD) model contains three main modules, i.e. Input block, the dual-DCNN unit, as shown in Fig. 2 and Multi-task learning component. The MD architecture takes a pair of images that are arbitrarily selected from the training data. The Dual DCNN contain DCNN-A and DCNN-B, which is the main two sequence learning module. The DCNN-A is pre-trained with the VGG16 network, and DCNN-B is pre-trained with ResNet50, and also both are fine-tuned to monitor the correct input sequence labels. The input block contains a pair of images where multi-scale dilated convolutional layers (MsDc) strategy is applied to both pairs. Then these pairs of images individually enter into DCNN-A and DCNN-B. A multi-task learning component is not only used to predict whether these pairs belong to the same class or not, but it also classifies the image pairs between benign and malignant. Furthermore, this component also helps to provide positive guidance if there is a synergistic mistake from both DCNNs.

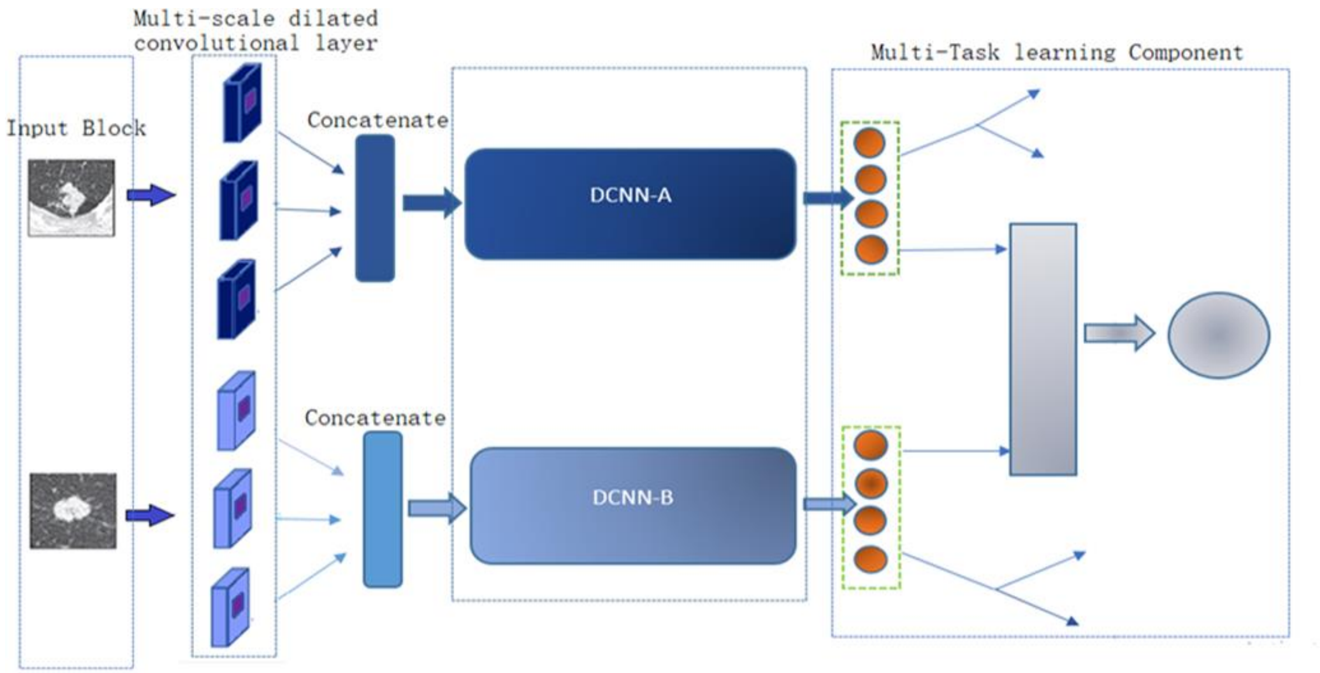

Multi-scale dilated convolutional layer
Multi-Task learning Component
Input Block
Concatenate
DCNN-A
Concatenate
DCNN-B

Fig 2. Architecture of Proposed MD Model.

B. Input Block and Multi-scale Dilated Convolutional Layer

Not quite the same as the traditional DCNNs, the suggested model randomly selects the MsDc pair of images as input from training data. Each image has its class label, and then it transfers to each DCNN. Each image is redimensioned to 224 × 224 × 1 by using a bicubic interpolation algorithm to unify the image size. Inside the input block, there are pair of grayscale images as input. The MsDc strategy is applied to the paired images before getting into DCNN (A/B). After employing this technique, the 3 channels image with different scale dilated convolution operation. Then 3 one channel images got from different receptive fields to one three-channel image is concatenated as the input of DCNN (A/B) as presented in Fig. 3. This approach uses different scale dilated filters to integrate the multi-scale contextual information systematically by extending the receptive area of the convolutional layers. The logic behind this approach is focused on the phenomenon that the dilated convolution could effectively extend the correct receptive area while maintaining useful contextual information. The MsDc composed of different dilation rates, which leads to various receptive filed within the input images. So, the dilation rate 3, 5, and 7 is used in MsDc. In Fig. 4, it is clearly shown how dilation allows the transfer of information. For the classification process to recognize this as a “Lung cancer” image, the circle region in the green cycle is most discriminating. To learn the corresponding feature representation at the area shown by the red cycle, a 3 ×3 convolutional kernel is implemented.

C. The Dual Deep Convolutional Nueral Network

Within this suggested model, the dual DCNN is an essential component with two full training units, DCNN-A and DCNN-B. Although an arbitrary structure of any DCNN such as GoogleNet or AlexNet can be implemented as a DCNN part in the recommended method. Both DCNNs is trained using X = {x(1), x(2)…. x(M)} image sequence and the corresponding Y = {y(1), y(2)…. y(M)} label class. For DCNN-A, a pre-trained VGG-16 model is used, which has thirteen convolutional layers and three fully-connected (FC) layers. Then a pre-trained ResNet 50 is adopted to initialize the DCNN-B because of the high representation capability of the popular residual network. It is concocted of 50 learnable layers, and it also trained for classification tasks on ImageNet datasets. In order to adopt the above models to our dataset, all fully connected (FC) layer is supplant to FC of 1024 neurons for DCNN(A/B) and then fine-tuning the ResNet-50 and VGG16 parameters by utilization our own training data. During optimization, the parameters of DCNN (A/B) are not shared mutually, which is denoted by θ A, and θ B. The uniform U (-0.05, 0.05) distribution of the weights of new FC layers is initialize. Both DCNNs is defined as the cross-entropy loss function as:

$$\iota(\theta) = \frac{1}{Q}\sum_{i=1}^{Q}[\log(\sum_{j=1}^{k} e^{\theta_j^t x(i)}) \text{-} \theta_{y(i)}^t\ x^{(i)}] \qquad (1)$$

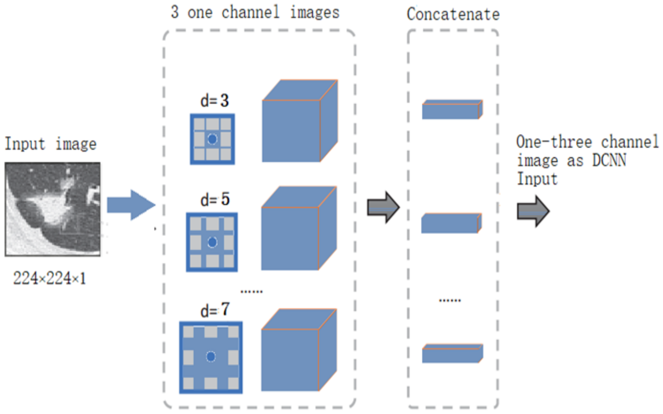


Fig 3. Structure of Multi-Scale Dilated Convolutional Layer Block.

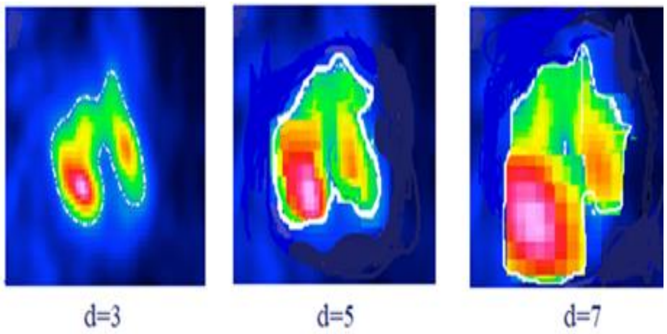


Fig 4. Reason for this Approach: the information can be moved through changing Dilated Levels from an Originally biased Area to other Regions.

Where Q is the number of training data. The Adam optimizer(mini-batch Adam) is used to optimize the θ. Input from a pair of images is accepted by both DCNN, which seek to control the training process with the true labels of the respective input sequence in each learning unit. While both DCNN has the ability to determine the input image label class, so, a multi-task learning component is produced that breaks the learning independency of the dual DCNNs by integrating activations from the last two FC layers in both DCNNs.

D. Multi-task Learning Component

To further track the training of each pair of images of each DCNN part, a multi-task learning component (MTLc) is designed, which consists of fully connected layers and vector concatenation layer, as shown in Fig. 5. Let the DCNNs components (A/B) have an image pair (xc, xd), respectively. The se` cond to last fc layer performance in the DCNN is described as the deep image features learned from that DCNN that can be accomplished through forwarding computation, as shown formally:

$$\mathrm{fc} = \mathrm{F}\big(\mathrm{xc}, \theta(\mathrm{i})\big) \tag{2}$$

$$\mathrm{fd} = \mathrm{F}(\mathrm{xd}, \theta(\mathrm{i})) \tag{3}$$

In the training data, the image pairs are arbitrary selected and denoted the attribute of a pair (xc, xd) as:

$$\mathrm{B}\ (\mathrm{xc}, \mathrm{xd}) = \begin{cases} 1 & \text{if } \mathrm{yc} = \mathrm{yd} \\ 0 & \text{if } \mathrm{yc} \neq \mathrm{yd} \end{cases} \tag{4}$$

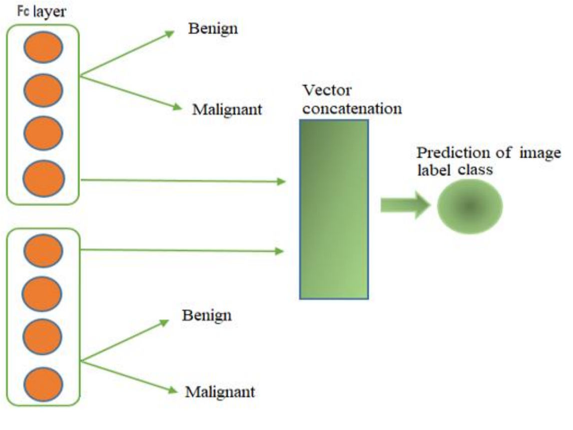


Fig 5. Diagram of Multi-Task Learning Component.

Where x_c and x_d are lung nodule images, y_c and y_d, respectively represent the true labels of x_c and x_d (like benign or malignant). The number of positive pairs, which is S=1 in the LIDC datasets is approximately 45% - 55% to prevent the unbalance data issue. After the output got from both DCNN A/B, two 1024 dimensional FC vectors, we copy and break them into 2 pieces for the multi-task learning part. In order to improve classification across learning tasks, the aim of multi-task is to accomplish mutual training by leveraging dependencies in the functionality in order to improve the performance of one task using the other. Moreover, softmax function is used as the non-linear activation function for the final prediction layer and use ReLU function as the non-linear activation for other fully connected layers. For each multi-task component, the hidden layers have 1000 neurons to solve the more complex problem. Furthermore, by using MTLc technique the two tasks is intended for prediction. The first task distinguishes among malignant and benign tumor while the other task will concatenate the vector from another DCNN network to predict the image label class. The analysis of the organized signal is useful to expedient the following binary entropy loss of the MTLc.

IV. Experiment

This section includes details of the implementation and analysis of the proposed model as well as a comparison of the experimental results. Section "A and B" illustrates how the hyperparameters are initiated and it also describes the dataset used in experiments. Section "C" presents the experiment on the LIDC dataset. The impact of λ is mainly tested on different values and compares the classification accuracy of the proposed model, and it also describes the comparison of MD model with several recently effective methods.

A. Dataset and Hyperparameter Setting

The LIDC dataset consists of 1018 CT scan patients of pulmonary cancer, and these CT images are divided into five categories. Detailed information for each nodule (diameter, coordinate, malignancy, texture, etc.) is indicated by 4 professional radiologists. The diameter of the nodules is between 3 mm and 30 mm. The interpolation spline technique is used as a method to separate with 1 mm × 1 mm due to the difference in resolution. In this study, the presumption is examined in the malignancy of nodules. Each nodule has an annotated malignancy by radiologists ranking from 1 to 5. As the final decision, the voting strategy is considered. If the nodule value was over 3 annotated by more than two radiologists, the nodule is considered as malignant. In contrast, the nodule is deemed benign. There are approximately 195 malignant nodules and 158 benign nodules. The nodules with the same votes are rejected. The central transect for each nodule voxel is extracted to minimize computational complexity. To alleviate the deep learning model overfitting, the Data Augmentation (DA) method is applied to enlarge the data by adding variants to a dataset. In particular, by zooming 0.2, the image is randomly flip and magnify. The selection of the translation step is from [-6, 6] voxels, and the angle of rotation was selected randomly from [90°, 180°, and 270°]. Finally, there are 1956 malignant nodules, 1862 benign nodules.

B. Hyperparameter Setting

For hyperparameter setting the number of iteration steps is set to 12000, and the learning rate is modified using an exponential decay process, and the initialize learning rate r =0.0002, with the learning rate attenuation of 0.95. The mini-batch Adam was acknowledged as an optimizer with a batch size 64. To stop the training procedure when the model is overfitting, 20% of training data were randomly selected to form a validation set, which was used to monitor the performance of prospective model.

As the training model, Keras is used to distinguish the benign and malignant lung nodules and Tensorflow as Keras backup. Python versions 2.7–3.6 are available in Keras. And it offers the following benefits: strong modularity, reduced simplicity, and scalability.

C. Experimental Result and Analysis

To evaluate the MD model performance on the LIDC dataset, first, the efficiency of the MD model is assessed y taking different values from the hyperparameter λ. The results in Fig. 6 indicate the best performance of suggested model with λ= 15. Then the classification accuracy of the proposed MD model is compared with the same experimental setting against the VGG-16, ResNet-50 model, and the proposed MD model is also evaluated without multi-scale dilated convolutional layer. The classification accuracies of the above models in each DCNN (A/B) are displayed in Fig. 7, and it is clearly shown that the recommended model attains the highest accuracy of 90.73% on the validation set. This reveals that each component of designed model, which is also VGG-16, ResNet-50 achieves an improvement in the precision of over more than 3% relative to the standard VGG-16 and ResNet-50 norm since integrating the multi-task learning component into a dual-DCNN architecture.

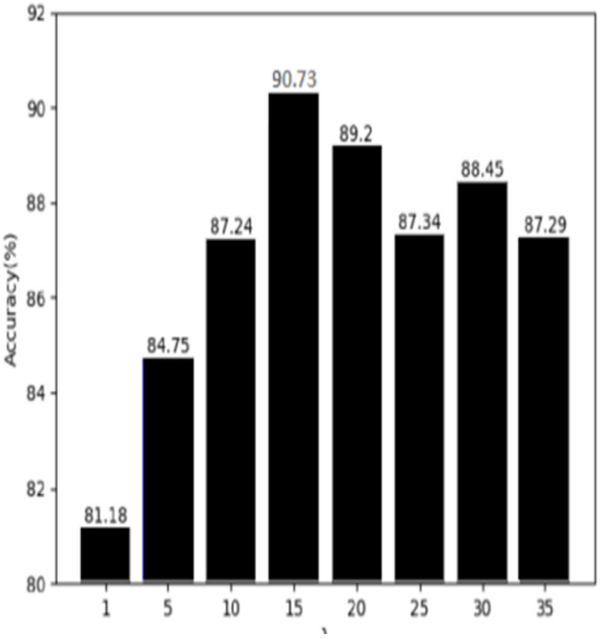


Fig 6. MD Model Performance with different λ Values.

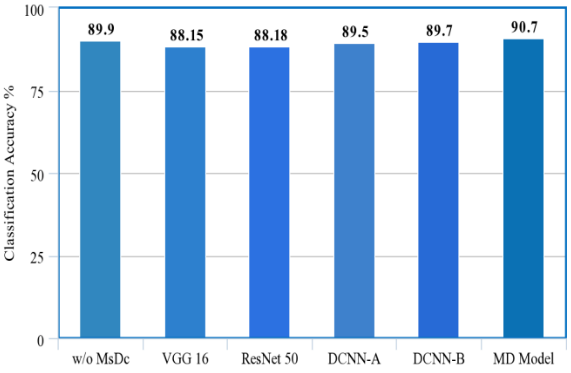


Fig 7. Classification Accuracy of Proposed MD Model w/o MsDc, with VGG 16, ResNet 50 , Dual DCNNs on the Validation set.

To assess the designed model's performance against the related methods, it is also compared to certain other effective methods with recent good results shown in Table I. Kumar et al. [29, 30] introduced a new ensemble system for the classification of medical images. The ensemble method has used many advanced CNNs as tailored extractors for image features that have captured the different information in medical images of different types. Zhang J et al. [31] have suggested a dual deep-convolutional neural network that is equipped with a synergic signal network to learn the representation of the image jointly and as well as a synergic signal system helps to verify the pair of the image belongs to the same category or not. Yuhai et al. [32] used an ensemble of 5 pretrained VGG and ResNet- 50 models, and 5 completely trained DCNN models, the ImageCLEF-2013 data set is augmented, and the baseline ResNet-50 model is improved by far. A hierarchical learning mechanism called Multi-scale Convolutional Neural Network (MCNN) [33] which is used to explore nodule complexity by eliminating discriminatory features from alternating stacked layers. The architecture uses multi-sized nodule patches to learn a collection of class features at the same time through concatenating response neuron activations of each input level at the last layer. For the classification of lung nodule [34], Haralick, Gabor, and LBP (local binary patterns) features are extracted and also implemented SVM classifier. They achieved 89.5% sensitivity and 86.02% specificity, respectively. Dhara et al. [35] based on the edge, shape, and textures features to ensure a benign and malignant classification.

However, the manual feature set is insufficiently suited as to whether the difference between the various pulmonary nodules types can be described accurately. Devinder Kumar et al. [36] suggested a CAD method utilizing deep functions separated from an auto-encoder for classifying pulmonary nodules as malignant or benign. In the analysis of 9 semantic features in CT images for lung nodules, Chen et al. [37] utilized three multi-task learning (MTL) systems for the use of various computational features extracted from deep learning systems of the convolutional neural network (CNN) and stacked denoising auto-encoder.

TABLE I. PERFORMANCE COMPARISONS WITH OTHER STATE-OF-ART METHODS FOR LUNG NODULE CLASSIFICATION TASK

Methods	Sensitivity %	Specificity %	Accuracy %
Davinder K et al. [36]	83.3	-	75.0
Kumar et al.[30]	-	-	77.5
Kumar et al.[29]	-	-	82.48
Zhang et al. [31]	-	-	86.58
Chen et al. [37]	60.3	95.4	86.8
Shen et al. [33]	-	-	86.84
Shen et al. [38]	77.0	93.0	87.1
Yuhai et al. [32]	-	-	87.37
Han et al. [34]	89.3	86.02	-
Dhara et al. [35]	89.7	86.3	-
Proposed MD Model	90.67	90.80	90.73

V. CONCLUSION

In this paper, a Multi-deep model is proposed for lung nodule classification and to resolve the problem posed by intra-class variation and inter-class similarity. In the first step, the MsDc method is applied to pair of images, which could help to increase the performance of lung nodule classification afterward our technique uses dual DCNNs with a multi-task learning component to allow dual DCNNs to learn from one another. It promotes the ability of the suggested model to distinguish between interclass samples that are easily ignored and the clear diversity of intra-class samples. The experimental result on the LIDC-IDRI dataset demonstrates that the proposed model attains the state-of-the-art performance in the lung nodule classification problem.

REFERENCES

[1] Zhao, B., Gamsu, G., Ginsberg, M. S., Jiang, L., & Schwartz, L. H. (2003). Automatic detection of small lung nodules on CT utilizing a local density maximum algorithm. journal of applied clinical medical physics, 4(3), 248-260.

[2] Valente, I. R. S., Cortez, P. C., Neto, E. C., Soares, J. M., de Albuquerque, V. H. C., & Tavares, J. M. R. (2016). Automatic 3D pulmonary nodule detection in CT images: a survey. Computer methods and programs in biomedicine, 124, 91-107.

[3] Lazebnik, S., Schmid, C., & Ponce, J. (2006, June). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) (Vol. 2, pp. 2169-2178). IEEE

[4] Fei-Fei, L., & Perona, P. (2005, June). A bayesian hierarchical model for learning natural scene categories. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) (Vol. 2, pp. 524-531). IEEE.

[5] Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., & Gong, Y. (2010, June). Locality-constrained linear coding for image classification. In 2010 IEEE computer society conference on computer vision and pattern recognition (pp. 3360-3367). IEEE.

[6] Wang, Z., Hu, Y., & Chia, L. T. (2013). Learning image-to-class distance metric for image classification. ACM Transactions on Intelligent Systems and Technology (TIST), 4(2), 1-22.

[7] Yang, Y., Yang, L., Wu, G., & Li, S. (2012, October). A bag-of-objects retrieval model for web image search. In Proceedings of the 20th ACM international conference on Multimedia (pp. 49-58).

[8] Sirinukunwattana, K., Raza, S. E. A., Tsang, Y. W., Snead, D. R., Cree, I. A., & Rajpoot, N. M. (2016). Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE transactions on medical imaging, 35(5), 1196-1206.

[9] Xie, F., Fan, H., Li, Y., Jiang, Z., Meng, R., & Bovik, A. (2016). Melanoma classification on dermoscopy images using a neural network ensemble model. IEEE transactions on medical imaging, 36(3), 849-858.

[10] Li, R., Zeng, T., Peng, H., & Ji, S. (2017). Deep learning segmentation of optical microscopy images improves 3-D neuron reconstruction. IEEE transactions on medical imaging, 36(7), 1533-1541.

[11] Chen, H., Qi, X., Yu, L., Dou, Q., Qin, J., & Heng, P. A. (2017). DCAN: Deep contour-aware networks for object instance segmentation from histology images. Medical image analysis, 36, 135-146.

[12] Jung, H., Kim, B., Lee, I., Lee, J., & Kang, J. (2018). Classification of lung nodules in CT scans using three-dimensional deep convolutional neural networks with a checkpoint ensemble method. BMC medical imaging, 18(1), 48.

[13] Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M., Eric, I., & Chang, C. (2014, May). Deep learning of feature representation with multiple instance learning for medical image analysis. In 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 1626-1630). IEEE.

[14] Li, X., Kao, Y., Shen, W., Li, X., & Xie, G. (2017, March). Lung nodule malignancy prediction using multi-task convolutional neural network. In Medical Imaging 2017: Computer-Aided Diagnosis (Vol. 10134, p. 1013424). International Society for Optics and Photonics.

[15] Wu, Z., Chen, J., Wang, Z., Su, J., & Cai, G. (2019, November). Multi-scale Convolutional Neural Network Based on 3D Context Fusion for Lesion Detection. In Chinese Conference on Pattern Recognition and Computer Vision (PRCV) (pp. 573-585). Springer, Cham.

[16] Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1717-1724).

[17] Mettes, P., Koelma, D. C., & Snoek, C. G. (2016, June). The imagenet shuffle: Reorganized pre-training for video event detection. In Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (pp. 175-182).

[18] Song, Y., Cai, W., Huang, H., Zhou, Y., Feng, D. D., Wang, Y., ... & Chen, M. (2015). Large margin local estimate with applications to medical image classification. IEEE transactions on medical imaging, 34(6), 1362-1377.

[19] Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530.

[20] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[21] Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., & Pande, V. (2015). Massively multitask networks for drug discovery. arXiv preprint arXiv:1502.02072.

[22] Deng, L., Hinton, G., & Kingsbury, B. (2013, May). New types of deep neural network learning for speech recognition and related applications: An overview. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 8599-8603). IEEE.

[23] Collobert, R., & Weston, J. (2008, July). A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning (pp. 160-167).

[24] Meng, Q., Bian, D., Guo, M., Lu, F., & Liu, D. (2012). Improved multi-scale retinex algorithm for medical image enhancement. In Information Engineering and Applications (pp. 930-937). Springer, London.

[25] Wang, Y., Wang, G., Chen, C., & Pan, Z. (2019). Multi-scale dilated convolution of convolutional neural network for image denoising. Multimedia Tools and Applications, 78(14), 19945-19960.

[26] Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., & Huang, T. S. (2018). Revisiting dilated convolution: A simple approach for weakly-and semi-

supervised semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7268-7277).

[27] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062.

[28] Gao, Y., Zhao, L., Wu, L., Ye, Y., Xiong, H., & Yang, C. (2019, July). Incomplete Label Multi-Task Deep Learning for Spatio-Temporal Event Subtype Forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 3638-3646).Kumar, A., Lyndon, D., Kim, J., & Feng, D. (2016, September). Subfigure and Multi-Label Classification using a Fine-Tuned Convolutional Neural Network. In CLEF (Working Notes) (pp. 318-321).

[29] Kumar, A., Kim, J., Lyndon, D., Fulham, M., & Feng, D. (2016). An ensemble of fine-tuned convolutional neural networks for medical image classification. IEEE journal of biomedical and health informatics, 21(1), 31-40.

[30] Kumar, A., Lyndon, D., Kim, J., & Feng, D. (2016, September). Subfigure and Multi-Label Classification using a Fine-Tuned Convolutional Neural Network. In CLEF (Working Notes) (pp. 318-321).s

[31] Zhang, J., Xia, Y., Wu, Q., & Xie, Y. (2017). Classification of medical images and illustrations in the biomedical literature using synergic deep learning. arXiv preprint arXiv:1706.09092.

[32] Yu, Y., Lin, H., Meng, J., Wei, X., Guo, H., & Zhao, Z. (2017). Deep transfer learning for modality classification of medical images. Information, 8(3), 91.

[33] Shen, W., Zhou, M., Yang, F., Yang, C., & Tian, J. (2015, June). Multi-scale convolutional neural networks for lung nodule classification. In International Conference on Information Processing in Medical Imaging (pp. 588-599). Springer, Cham.

[34] Han, F., Wang, H., Zhang, G., Han, H., Song, B., Li, L., ... & Liang, Z. (2015). Texture feature analysis for computer-aided diagnosis on pulmonary nodules. Journal of digital imaging, 28(1), 99-115.

[35] Dhara, A. K., Mukhopadhyay, S., Dutta, A., Garg, M., & Khandelwal, N. (2016). A combination of shape and texture features for classification of pulmonary nodules in lung CT images. Journal of digital imaging, 29(4), 466-475.

[36] Kumar, D., Wong, A., & Clausi, D. A. (2015, June). Lung nodule classification using deep features in CT images. In 2015 12th Conference on Computer and Robot Vision (pp. 133-138). IEEE.

[37] Chen, S., Qin, J., Ji, X., Lei, B., Wang, T., Ni, D., & Cheng, J. Z. (2016). Automatic scoring of multiple semantic attributes with multi-task feature leverage: A study on pulmonary nodules in CT images. IEEE transactions on medical imaging, 36(3), 802-814.

[38] Shen, W., Zhou, M., Yang, F., Yu, D., Dong, D., Yang, C., & Tian, J. (2017). Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. Pattern Recognition, 61, 663-673.

ECG and EEG Pattern Classifications and Dimensionality Reduction with Laplacian Eigenmaps

Monica Fira[1]
Institute of Computer Science
Romanian Academy, Iasi, Romania

Liviu Goras[2]
Institute of Computer Science, Romanian Academy
"Gheorghe Asachi" Technical University of Iasi, Romania

***Abstract*—In this paper, we investigate the effect of dimensionality reduction using Laplacian Eigenmap (LE) in the case of several classes of electroencephalogram (EEG) and electrocardiographic (ECG) signals. Classification results based on a boosting method for EEG signals exhibiting P300 wave and k-nearest neighbour for ECG signals belonging to 8 classes are computed and compared. For EEG signals, the difference between the rate of classification in the original and reduced space with LE is relatively small, only several percent (maximum 10% for the 3 – dimensional space), and the original EEG signals belonging to a 128-dimensional space. This means that, for classification purposes the dimensionality of EEG signals can be reduced without significantly affecting the global and local arrangement of data. Moreover, for EEG signals that are collected at high frequencies, a first stage of data preprocessing can be done by reducing the dimensionality. For ECG signals, for segmentation with and without centering of the R wave, there is a slight decrease in the classification rate at small data sizes. It is found that for an initial dimensionality of 301 the size of the signals can be reduced to 30 without significantly affecting the classification rate. Below this dimension there is a decrease of the classification rate but still the results are very good even for very small dimensions, such as 3. It has been found that the classification results in the reduced space are remarkable close to those obtained for the initial spaces even for small dimensions.**

Keywords*—Laplacian Eigenmaps; dimensionality reduction; biosignals; electrocardiographic signal (ECG); electroencephalogram (EEG)

I. Introduction

Manifold learning is a class of methods aimed at evidencing low-dimensional manifolds embedded in a high-dimensional ambient space. The concept is closely related to dimensionality reduction according to the assumption that for high dimensional spaces, the data is expected to "live" in a (much) lower dimensional space or, in the nonlinear case, on a (much) lower dimensional manifold. In other words, whether linear manifold learning does not result in a good low-dimensional representation of high-dimensional data, it might happen that data lie on or close a nonlinear manifold so that more powerful non-linear dimensionality reduction by preserving the local structure of the input data can be applied. If data stay on a low-dimensional nonlinear manifold, it has been shown that usual methods will adjust automatically, and better learning rates may be obtained even if one understands little about the manifold form [1-4]. However, even when it is known that data are on a nonlinear manifold there are circumstances when the algorithms fail to recover the manifold [5]. Starting from the above considerations regarding the nature of signals, manifolds and supervised learning, we asked the question that if for a class of real data we can reduce the size of the signals and if a supervised classification obtains similar results on the real, original data space and on the reduced space [6].

In last years, manifold learning methods have grown explosively [17-19]. A classification from the point of view of preserving the geometry, the methods of manifold learning can be classified into two broad categories, namely:

a) Methods with preserving the local geometry structure: locally linear embedding (LLE) [7], Laplacian eigenmaps (LE) [1], manifold charting (MC) [8], Hessian locally linear embedding (HLLE) [9].

b) Methods with preserve the global characteristics: isometric mapping (ISOMAP) [10], diffusion map [11]

The LE algorithm has been initially applied on real signals in the medical field. Without a thorough analysis, in 2007 it was tested by Gramfortin and Clerc [12] on MRI images and signed EEG. Lashgari and Demircan in 2017 [13] used the LE algorithm in Electromyography (EMG) signal classification problems.

For medical signals such as ECG and EEG, in 2016 Erem et al. [14] presents the Laplacian Eigenmaps machine learning algorithm combined with dynamical systems ideas for analyze emerging dynamic behaviours.

The method chosen in this paper for dimensionality reduction of electroencephalogram (EEG) and electrocardiographic (ECG) signals is the Laplacian Eigenmap [1]. The outcomes reported here extend our previous results published in [15 - 16], where the performances of the LE algorithm were tested only on ECG time signals and where a comparative analysis between the LE and LPP (Locality Preserving Projections) algorithms was done. Here we propose a more rigorous analysis of the results obtained with LE for both ECG and EEG signals. These two classes of signals were chosen since they are also the most used 1D signal in the field of bio signal processing.

In order to evaluate the effect of dimensionality reduction in both cases, EEG and ECG, we compare the classification rates obtained with the original data with those obtained on the EEG and ECG segments on which various degree of dimensionality reduction were obtained using Laplacian Eigenmaps (LE).

Next, we will analyze the effect of reducing the dimensionality of the data. For this we will calculate the classification rate in the initial space and the classification rate in the reduced space. If the two classification rates are close, it means that close neighbours remain close, meaning the geometry is preserved, at least the local geometry. For this we will use two types of signals, namely, ECG and EEG signals. For each signal type we will choose a classification problem specific to this one with which we have worked and we have obtained good results. Then we will reduce the dimensionality of the signals and using the same classifier we will compare the classification rates obtained in the initial space and those obtained in the reduced space. In Section II the theoretical part of the Laplacian Eigenmaps algorithm is presented, in Section III we will present the segmentation method and the classifier chosen for EEG signals (EEG signal acquired by Hoffmann and collaborators in their laboratory and the Gradient boosting classifier) and for type signals. ECG (MIT-BIH Arrhythmia database and segmentation with / without R wave centring and a KNN classifier with Euclidean distance and the nearest neighbour membership decision).

II. Laplacian Eigenmaps

The target of the LE algorithm is to find a low-dimensional data representation but to conserve the local geometry of the data. This preservation of the geometry is based on the distances between the pairs of near neighbours on the manifold.

The LE algorithm associates the data with a graph with weights. These weights are calculated based on the distances between neighbours. The weights thus found are used to minimize a cost function that finds a mapping from the initial data to a small dimensional space [1] [13-14].

The explanation of the weights calculated based on the neighbourhoods is that the distance in the low-dimensional data representation between a data point and its first nearest neighbour contributes more to the cost function compared to the distance between the data point and its second or the other nearest neighbour. The minimization of the cost function is defined as an eigenproblem [6].

The LE algorithm [1] construct a neighbourhood graph G in which every data point xi is connected by an edge to its k nearest neighbours. In our case, for all points xi and xj in G that are connected by an edge within a neighbourhood Ni, a weight is computed using the Gaussian kernel function,

$$\mathrm{w_{ij}} = \mathrm{w}(\mathrm{x_i} - \mathrm{x_j}) = \begin{cases} \exp\left\{-\dfrac{\|\mathrm{x_i} - \mathrm{x_{ij}}\|^2}{2\sigma^2}\right\}, & if\ x \in \mathrm{N_i}; \\ 0, & otherwise. \end{cases}$$

where σ is a constant called heat kernel parameter, leading to a sparse matrix W that is symmetric adjacency. It is desired that points x_i, x_j that are close to the initial spatial map are mapped to points y_i, y_j to remain close and in the small space. This can be achieved by minimizing the cost function

$$\emptyset(\mathrm{Y}) = \sum_{\mathrm{ij}} \|\mathrm{y_i} - \mathrm{y_j}\|^2 \mathrm{w_{ij}}$$

where large weights w_{ij} correspond to small distances between the high-dimensional data points x_i and x_j. Hence, the difference between their low-dimensional representations y_i and y_j highly contributes to the cost function. As a result, the close points of the high-dimensional space are placed as close as possible in the low-dimensional space [1-2].

Then follows the last stage of the LE algorithm, namely, the calculation of eigenvalues and eigenvalues for the general eigenvector problem,

$$\mathrm{Lf} = \lambda \mathrm{Df}, \tag{1}$$

where D = (d_{ij}) is an (n×n) diagonal matrix with elements

$$\mathrm{d_{ii}} = \sum_{\mathrm{j \in N_i}} \mathrm{w_{ij}}$$

and matrix L is calculated based on matrices G and D, namely, L = D − W is the Laplacian matrix which is symmetric and positive semidefinite. The L matrix can be thought of as an operator on functions defined on the vertices of G.

Mapping in the low-dimensional space is done by eliminating the eigenvector f_0 corresponding to eigenvalue 0 and using the next m eigenvectors corresponding to the next eigenvalue. The embedding in an m dimensional Euclidean space is:

$$x_i \rightarrow (f_1(i), \ldots, f_m(i)).$$

where $f_0, \ldots, f_{k-1}$ are the solutions of equation (1), in ascending order of their eigenvalues [1].

III. Experimental Results

In what follows we will present several classification results for EEG and ECG signals seen only as a measure of the conservation of the spatial geometry on manifolds and not of the quality of the classifier. In other words we will use the classification rate as a measure of preserving geometry, i.e. find how much the classification rate decreases when reducing the space dimension with the LE algorithm.

A. EEG Signals

Starting from the results obtained in our paper [20], in which we used the EEG signal to verify the preservation of the neighbourhoods in the reduced space with compressed sensed (CS), using the same test data we check if the reduced dimensionality data with LE keeps its neighbours.

In paper [20] we used compressed sensed algorithm to reduce the EEG data size. The common point of the paper [22] with the present paper is that the same EEG data is used to test the methods (in fact the same EEG database) and the same classifier, namely gradient boosting. The difference between these papers is that the method of decreasing the dimensionality of the data is distinct.

For testing the method there were used EEG signals acquired by Hoffmann and collaborators in their laboratory - a reduced database is available on the internet at [21]. The database includes EEG signals collected for 32 channels, which are grouped in 942 vectors for classification and lasting 1 sec each. The Gradient boosting classifier from [22] was used. It should be noted that the used software was developed by the authors as a machine learning method and creates a powerful algorithm from several poor classifiers.

In the above work, the authors described a simple and powerful method to detect the P300 from single EEG trials which have been used to build a P300 based spelling device for BCI. To compute from training data a function that detects P300s from single EEG trials, boosting has been used to stepwise maximize the Bernoulli log-likelihood of a logistic regression model.

We mention that we kept the configuration parameters for gradient boosting method were kept the same as in [22]. Thus the maximal number of iterations is Mmax = 200, the best M was 30×10 cross-validation loop, and ε = 0.05(same setting as in [22]). The results are presented in Table I and, in more detail, in Fig. 1 and Fig. 2.

Fig. 1 shows three EEG signals in the 128 dimensional space and their mapping on the space spanned by the first 30 eigenvectors. It happens that with the reduction of the spatial dimensions the signal waveforms change, but the relative distances are preserved as it will be illustrated in Fig. 3.

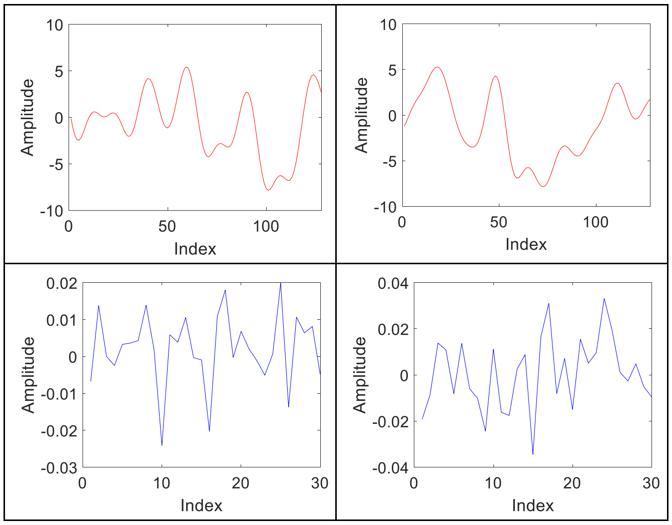


Fig. 1. EEG Signals in a 128 and 30 Dimensional Space (First/Red and Second/Blue, Respectively).

TABLE I. MAXIMUM CLASSIFICATION RATE FOR ORIGINAL EEG SIGNALS AND EEG SIGNALS WITH REDUCED SPACE (WITH LAPLACIAN EIGENMAPS ALGORITHM) FOR GRADIENT BOOSTING FOR SEVERAL CONFIGURATIONS OF CHANNELS

	23 channels (CP1, CP5, P3, Pz, PO3, PO4, P4, C4, FC6, FC2, F4, AF4, Fp2, Fz, Cz, Fp1, AF3, F3, Fc1, Fc5, C3, CP6, CP2)	**8 channels** (Fz,Cz,Pz,Oz, P7, P3, P4, P8)	**4 channels** (Fz,Cz,Pz, Oz)
Original EEGs (128 dimensional space)	86%	85%	80%
3 D	78%	80%	73%
5 D	79%	81%	76%
15 D	83%	82%	77%
30 D	84%	83%	79%

Table I shows a very small difference in the classification between original EEGs - 128 dimensional space and EEG with 30 - dimensional space - the classification rate decreases by only 1 or 2 percent. The decrease of the classification rate from 30 to 15 dimensional space is also about 1%.

Fig. 2 show the accuracy obtained after the cross-validation loop for configurations with 23, 8 and 4 channels for original EEGs and for 3, 5 or 15-dimensional spaces obtained with LE. As it can be seen, the gradient boosting algorithm converges to an optimal solution. The difference between the rate of classification in the original and reduced space with LE is relatively small, only several percents (maximum 10% for the 3 – dimensional space), and the original EEG signals belonging to a 128-dimensional space.

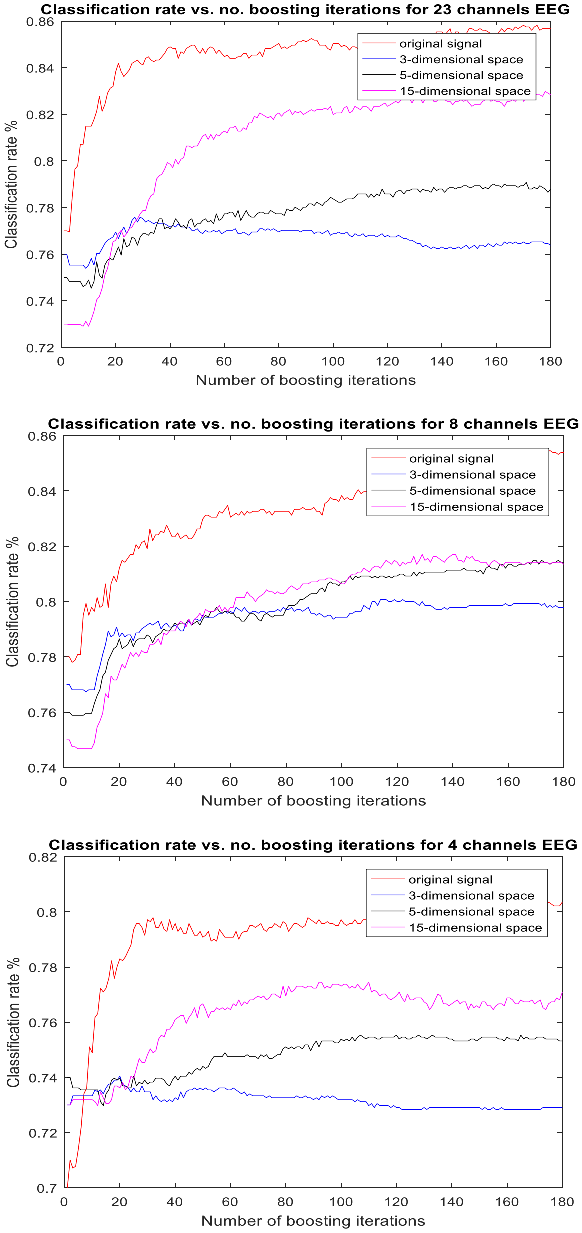


Fig. 2. Classification Performance for different Values of M for Several Space Dimensions for the Reduced EEG Signals (23, 8 and 4 Channels).

This result confirms that the global data structure is preserved and that a classification can be made in the small space with results very close to the classification in the initial space. This result is kept regardless of the channel configuration.

Remarkably, the classification rate decreases very little with the reduction of the signal space and that the trend of evolution according to the number of iterations is kept the same for all space dimension. Another interesting result is that the sigma parameter in the Gaussian distribution has almost no influence on classification rate performances as shown in Table II where it can be seen that the classification rate is slightly affected with the modification of sigma, the maximum difference being 3%.

TABLE II. INFLUENCE OF THE SIGMA PARAMETER ON THE CLASSIFICATION RATE FOR NEIGHBOURHOOD K = 5 AND 15- DIMENSIONAL SPACE

Sigma	Classification rate %
1	81%
4	81%
7	81%
10	81%
13	83%
16	82%
19	80%
22	80%
25	82%
28	82%

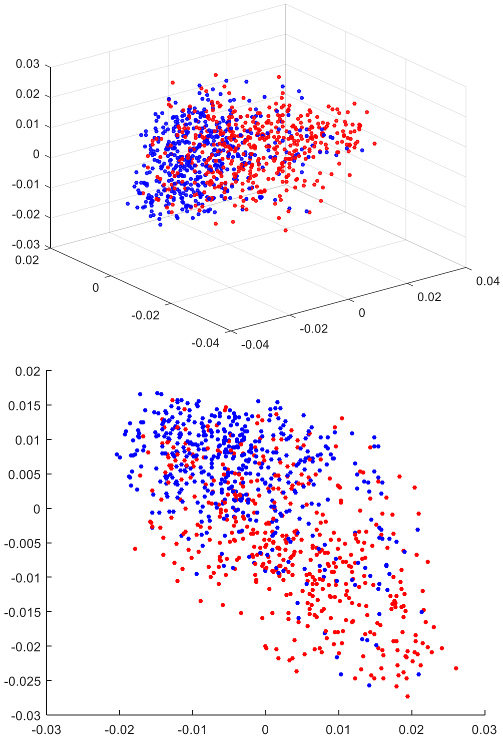


Fig. 3. EEG Data Mapped into a 3 Respectively 2-Dimensional Space for Sigma = 5, Nearest Neighbours k = 7 and a Classification Rate = 78,37% (Fz Channel).

To make an intuitive image on data, in Fig. 3 we present two examples of EEG signal data in reduced spaces with 3 and 2 dimensions using the LE algorithm.

B. ECG Signals

In the case of ECG signals, the starting point are the results presented in [15] where the results obtained with Laplacian Eigenmaps (LE) and Locality Preserving Projections(LPP) are analyzed and compared to reduce the dimensionality of the signal space. In [15] it was found that for small sizes LE offers better results. In this paper, we analyze whether the centring of the R wave brings significant improvements for very low-dimensional space (such as 2D and 3D).

For ECG signals, we have used 44 ECG from the MIT-BIH Arrhythmia database. The ECG signal was acquired at a sampling frequency of 360Hz, with 11 bits / sample [23]. In addition to the ECG signals, the database also comprises annotation files with the index of the R wave and the class for each ECG beats. In the database were identified 8 major classes of pathologies (from which 7 classes of pathological beats.

We used two different methods of segmenting ECG signals, namely:

- Segmentation with re-sampling (301 samples per signal)
- Segmentation with re-sampling as above and R waves centred.

a) Segmentation with re-sampling: A cardiac beat begins in the middle of the RR interval and ends in the middle of the next RR interval.

b) Segmentation with re-sampling and R waves centred: For the second splitting up method, to increase the classification rate we used the method reported in [24], namely, starting with ECG signals for which the position of the R-wave has been exactly determined. A cardiac beat begins in the middle of the RR interval and ends in the middle of the next RR interval as before and in the cardiac beats thus obtained, the R wave will be positioned in the middle by resampling the waveforms on both sides of R. In this way patterns with the centred cardiac R wave have been obtained. In this case, all cardiac patterns are of size 301 as before, the R wave being positioned on the 150th sample.

The database thus constructed contains 5608 patterns, each class having 700 such patterns (7 pathological and 1 normal). The results are presented in Table III and Fig. 4.

For classification, the KNN classifier with Euclidean distance and the membership decision was based on the nearest neighbour was used.

Table III shows a small difference in the classification rate of original cardiac patterns - 301 dimensional space and cardiac patterns with 30 - dimensional space. The classification rate decreases by approximately 3 percent. The decrease of the classification rate from 30 to 15 dimensional space is only 1%. These proportions are similar no matters if the R wave is centred or not.

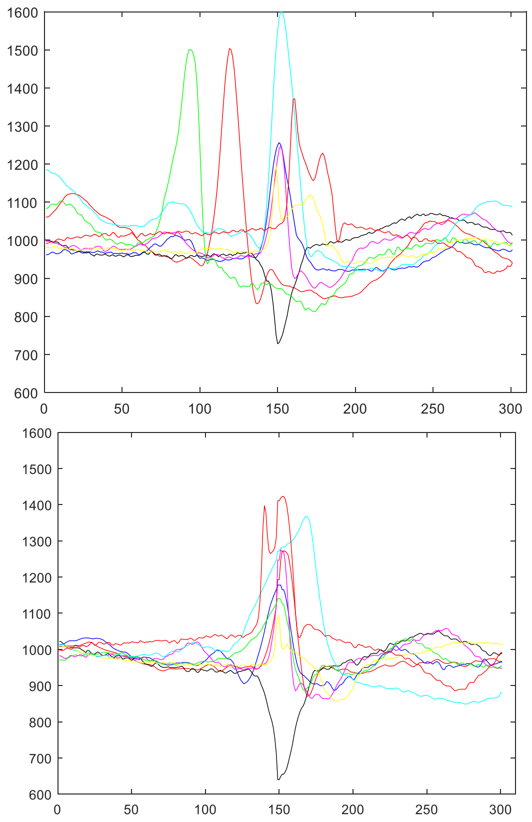


Fig. 4. Samples of Segmentation ECG Signals from each of the Eight Pattern Classes (up - Segmentation with re-Sampling; down - Segmentation with re-Sampling and R waves Centred).

TABLE III. CLASSIFICATION RATE FOR ORIGINAL ECG SIGNALS AND ECG SIGNALS IN REDUCED SPACES

	Segmentation with R centred	**Segmentation without R centred**
ECG originals (301 dimensional space)	92.33%	90%
2 - D	75,67%	72,56%
3 - D	85,27%	83,61%
5 - D	86,85%	85,36%
15 - D	88,32%	86,50%
30 - D	89,32%	86,97%

In Fig. 5, we present classification rates vs. space dimension for LE (for sigma = 5 and neighbourhood k = 9) for ECG segments without R wave centred (blue) and segmentation with R wave centred (red). It can be observed that there is a slight decrease in the classification rate for both original signals (in 301 dimensional space) and in the reduced space. Thus, for the original signals a 90.36% classification rate is obtained if there is no R wave centring compared to 92.5% for segmentation with centred R wave. In the above conditions, for the initial ECG signals the classification error for the 8 classes was found to be 2%, this small difference being significantly the result of the R-wave centring.

Because LE offers very good results for the very small size of the space, the method can be used for data represented in 2D or 3D to give us a visual idea of the spatial distribution of data in classes. This visualization can be very useful to understand the spatial arrangement of some data, an arrangement that can sometimes be very twisted and the choice of the classifier or some parameters of the classifier is related to the spatial arrangement of the data.

In Fig. 6(a and b) the ECG signal data in 3D (normal and zoom for the central zone) mapping are shown.

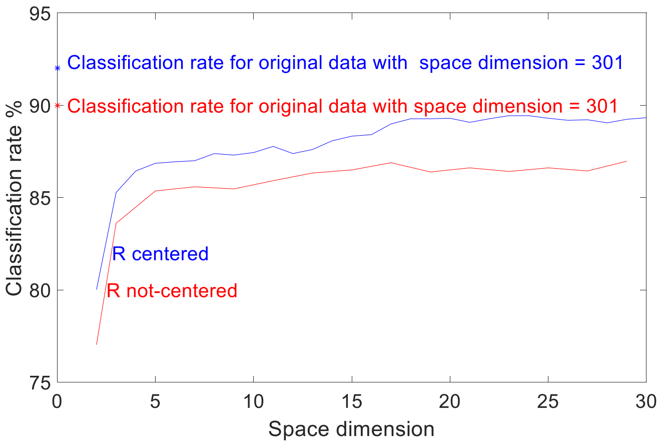


Fig. 5. Classification Rate vs. Space Dimension for Laplacian Eigenmaps (Sigma = 5, Neighbourhood k = 9) for Segmentation with Centred and not-Centred R wave (red = not-Centred and Blue = R Centred).

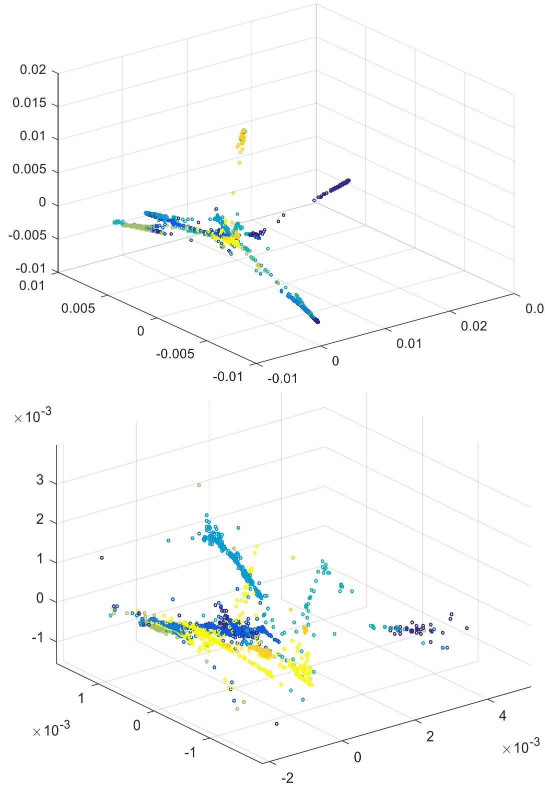


Fig. 6. ECG Data Mapped into a 3-Dimensional Space with LE for Sigma = 5, Nearest Neighbours k = 7 and a Classification Rate = 85% for Segmentation with R Centred (Left = 3D Plot and Right = Zoom from Central Region).

IV. Discussions

For EEG signals it has been found that the gradient boosting algorithm converges to an optimal solution. The difference between the rate of classification in the original and reduced space with LE is relatively small, only several percent (maximum 10% for the 3 – dimensional spaces), and the original EEG signals belonging to a 128-dimensional space. This means that, for classification purposes the dimensionality of EEG signals can be reduced without significantly affecting the global and local arrangement of data. Moreover, for EEG signals that are collected at high frequencies, a first stage of data pre-processing can be done by reducing the dimensionality. Another observation for EEG signals is that the classification rate decreases very little with the reduction of the signal space and that the trend of evolution according to the number of iterations is the same for all space dimensions. It is also observed that the sigma parameter in the Gaussian distribution has almost no influence on classification rate performances.

For ECG signals, for segmentation with and without centring of the R wave, there is a slight decrease in the classification rate at small data sizes. It is found that for an initial dimensionality of 301 the size of the signals can be reduced to 30 without significantly affecting the classification rate. Below this dimension there is a decrease of the classification rate but still the results are very good even for very small dimensions, such as 3 (classification rate decreases from 92.33%% for initial ECG signals with 301 dimensionality to 89.32% for dimensionality 3).

At present, we are not aware of studies similar to the application of the LE algorithm for both EEG and ECG time signals thus a comparison of our results obtained with this algorithm with other authors is not possible. However, below we present in Table IV with results reported in [25] with PCA, LDA, KPCA, Isomap and LE *only* for ECG signals. It can be observed in Table IV that our results with LE in spaces with reduced dimensionalities are similar with the observation that they were not obtained on the same database.

TABLE IV. Classification Rate for ECG Signals in Dimensions 5 and 10 Presented in [15] and our Results

	10 dimensions	**5 dimensions**
PCA [25]	75,25	75,83
LDA [25]	69,84	65,72
KPCA [25]	85,04	84,35
Isomap [25]	83,39	85,67
Laplacian Eigenmaps (LE) [25]	**88,04**	**86,69**
Our results with LE	**88,24**	**86,85%**

V. Conclusions

The remarkable result reported in this paper is the fact that dimensionality reduction for EEG and ECG signals using LE does not affect significantly the classification rate even for rather small dimensions. This proves not only that the neighbourhoods are preserved by LE but also that the signals have a significant robustness regarding classification when mapped on low dimensional manifolds. This allows having an intuitive image of the spatial distribution for the case of 2D or 3D when it is possible to plot the data.

In the future, we aim to use the advantages offered by the LE algorithm for classification problems and to find solutions for new data (i.e. so that we would not need a new recalculation whenever we have a new data).

References

[1] M. Belkin, and P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15:1373–1396 (2003).

[2] M.Belkin, P.Niyogi, and V. Sindhwani, Manifold regularization: a geometric framework for learning from labeled and unlabeled examples, Journal of Machine Learning Research, 7, 2399–2434, (2006)

[3] L.J.P. van der Maaten, E. O. Postma , H. J. van den Herik, Dimensionality Reduction: A Comparative Review (2008)

[4] Ma Yunqian and Yun Fu, Manifold Learning Theory and Applications, CRC Press, 2012

[5] Y. Goldberg, A. Zakai, D. Kushnir, Y. Ritov, Manifold learning: the price of normalization, Journal of Machine Learning Research, 9, 1909–1939., 2008.

[6] Samuel, Gerber & Tasdizen, Tolga & Fletcher, P. & Joshi, Sarang & Whitaker, Ross. (2010). Manifold Modeling for Brain Population Analysis. Medical image analysis. 14. 643-53. 10.1016/j.media.2010.05.008.

[7] S.T. Roweis, L.K. Saul, Nonlinear Dimensionality Reduction By LocallyLinear Embedding, Science, 290(2000)2323-2326.

[8] M. Brand, Charting a manifold, Proceedings of Neural Information Processing Systems, 2002.

[9] Z. Zhang,H. Zha,Principal Manifolds and Nonlinear Dimension Reduction via LocalTangent Space Alignment,SIAM J. Scientific Computing, 26(1)(2005)313-338

[10] J.B. Tenenbaum, V. de Silva, J.C. Langford, A Global GeometricFramework for Nonlinear Dimensionality Reduction,Science, 290(2000) 2319-2323.

[11] R.R.Coifman, S.Lafon, A.B.Lee,M.Maggioni, B. Nadler, F. Warner, S. W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps,PNAS, 102(2005)7426–7431

[12] Alexandre Gramfort, Maureen Clerc, Low Dimensional Representations of MEG/EEG Data Using Laplacian Eigenmaps, Joint Meeting of the 6th International Symposium on Noninvasive Functional Source Imaging of the Brain and Heart and the International Conference on Functional Biomedical Imaging, (2007)

[13] Elnaz Lashgari, Emel Demircan, Electromyography Pattern Classification with Laplacian Eigenmaps in Human Running, World Academy of Science, Engineering and Technology (2017).

[14] B. Erem, R. Orellana Martinez, D.E. Hyde, J.M. Peters, F.H. Duffy, P. Stovicek, S.K. Warfield, R.S. MacLeod, G. Tadmor, D.H. Brooks, Extensions to a manifold learning framework for time-series analysis on dynamic manifolds in bioelectric signals, Phys Rev E. Apr;93(4), (2016)

[15] M. Fira, L. Goras, "Dimensionality Reduction for ECG Signals; Laplacian Eigenmaps and Locality Preserving Projections", ISSCS 2019, Iasi (2019)

[16] M Fira, L. Goras, "On SomeMethods for DimensionalityReduction of ECGSignals", International Journal of Advanced Computer Science and Applications, Vol. 10, No. 9, (2019)

[17] Chia-Hung Wei, Yue Li, "Machine Learning Techniques for Adaptive Multimedia Retrieval: Technologies Applications and Perspectives", 10.4018/978-1-61692-859-9

[18] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment, "SIAM J. Sci. Comput., vol. 26, no. 1, pp. 313–338, 2005.

[19] Bo Li, Yan-Rui Li, Xiao-Long Zhang, "A Survey on Laplacian Eigenmaps Based Manifold Learning Methods, Neurocomputing", (2018), https://doi.org/10.1016/j.neucom.2018.06.077

[20] M. Fira, "The EEG Signal Classification In Compressed Sensing Space", 23 -27 iulie 2017, The Twelfth International Multi-Conference on Computing in the Global Information Technology, ICCGI 2017, Nice, Franta

[21] http://mmspg.epfl.ch/cms/page-58322.html (22 May 2017)

[22] U. Hoffmann, G. Garcia, J.-M. Vesin, K. Diserens, T. Ebrahimi, "A Boosting Approach to P300 Detection with Application to Brain-Computer Interfaces", Proceedings of IEEE EMBS Conference on Neural Engineering, (2005)

[23] http://www.physionet.org/physiobank/database/mitdb/

[24] M. Fira, L. Goras, N. Cleju, C. Barabasa „On the classification of compressed sensed signals", ISSCS 2011, Iasi (2011)

[25] Çiğdem Bakır, Classification of ECG Signals By the Neighborhood Feature Extraction Method, Balkan Journal Of Electrical & Computer Engineering, Vol. 3, No. 3 (2015)

A Solution to the Hyper Complex, Cross Domain Reality of Artificial Intelligence: The Hierarchy of AI

Dr Andrew Kear[1]
Faculty of Media and Communication
Bournemouth University
England

Sasha L. Folkes[2]
Marketing Intelligence
London

***Abstract*—Artificial Intelligence (AI) is an umbrella term used to describe machine-based forms of learning. This can encapsulate anything from Siri, Apple's smartphone-based assistant, to Tesla's autonomous vehicles (self-driving cars). At present, there are no set criteria to classify AI. The implications of which include public uncertainty, corporate scepticism, diminished confidence, insufficient funding and limited progress. Current substantial challenges exist with AI such as the use of combinationally large search space, prediction errors against ground truth values, the use of quantum error correction strategies. These are discussed in addition to fundamental data issues across collection, sample error and quality. The concept of cross realms and domains used to inform AI, is considered. Furthermore there is the issue of the confusing range of current AI labels. This paper aims to provide a more consistent form of classification, to be used by institutions and organisations alike, as they endeavour to make AI part of their practice. In turn, this seeks to promote transparency and increase trust. This has been done through primary research, including a panel of data scientists / experts in the field, and through a literature review on existing research. The authors propose a model solution in that of the Hierarchy of AI.**

Keywords—Artificial intelligence; classification; ground truth value; Hierarchy of AI; Model of AI

I. Introduction

A great deal of public funded investment is going into AI and yet the authors propose that there are still some fundamental issues with the different classifications and thus understanding of AI. These in turn lead to under confidence which could be overcome with more logical classification. The literature on AI is vast and multifaceted. Below, the authors have grouped together a series of definitions that vary according to the source:

AI can be defined as "any system . . . that generates adaptive behaviour to meet goals in a range of environments can be said to be intelligent" [23];

AI can be seen as "intelligent systems'' that "are expected to work, and work well in many different environments [27]. Their property of intelligence allows them to maximize the probability of success even if full knowledge of the situation is not available'';

AI is also defined as a division of computer science, in particular, "the study of the relation between computation and cognition'' [5];

Others [61] note how AI is a "big field'' that can be defined as "the study of agents that receive precepts from the environment and perform actions''.

In addition to the above there are sources that argue AI is the wrong term entirely:

Psychometric Artificial Intelligence (PAI) is according to [9] more suitable, since it refers to "building information-processing entities capable of at least solid performance on all established, validated tests of intelligence and mental ability'' (including artistic and literary creativity/ mechanical ability);

An alternative view [70] argues that the lack of consistency in definitions goes beyond semantic differences, as it poses a threat to developments in the field. With multiple definitions, "progress made under one characterization of AI is not viewed as success by others who operate under a different perception of it'' [70]. This results in diminished confidence in the field, as well as "promote(ing) premature conclusions of what can and cannot be accomplished and limit progress and funding along research paths''.

It is argued [45] that much of the trouble around defining AI is due to the issues with "intelligence''. In their 2007 paper, 70 definitions of "intelligence'' were gathered to display the disparity. An alternative view [2] highlights how, "curiously, the lack of a precise, universally accepted definition of AI probably has helped the field to grow, blossom, and advance"; several scholars argue for "changing the language used around AI to sharpen its conceptual clarity'' [43].

This is particularly relevant, as awareness of AI is rising, albeit with muddled meaning. In a nationally representative survey by [14] which interviewed 1078 respondents, just 42% were able to provide a credible definition of AI. A quarter of respondents described AI as relating to robots, and of all responses, the majority had the view that AI causes significant anxiety. This is problematic for industries such as eHealth (electronic health), as it is reliant on patient trust; similarly with schooling, public services and retail.

Despite several breakthroughs, AI is not without criticism. It is highlighted [79] that while accidents involving software or robotics "can be traced back to the early days of such technology'', AI failures "are directly related to the mistakes produced by the intelligence such systems are designed to exhibit''. Instead of learning an intended function, an AI system can adapt an alternate function to reach the predefined goal. In 2016, OpenAI trained an AI agent to play CoastRunners, an online multiplayer game. Instead of finishing the course, the AI agent found a way of winning the game without completing the course: by repeatedly turning around in a circle timed to coincide with reappearing targets. Though somewhat harmless in this context, the above example highlights the dangers of reinforcement learning using imperfect proxies [79].

In more practical settings, AI systems have been known to stereotypically assign professions to genders on LinkedIn and fail to recognise ethnic differences in smartphone ID systems, in the case of Apple’s iPhone X [11]. In extreme cases, AI failures have resulted in death: with Uber's self-driving car crashing into a pedestrian in 2018 as an example. More recently in the news, DeepFakes have seen celebrities faces imposed onto pornography, political speeches and more. Additionally, issues with AI determining which inmates get parole have also been highlighted and subsequently noted [72] where “its potential dangers are serious and far-reaching: if video evidence is no longer credible, this could further encourage the circulation of fake news’’. While the harmful effects of the above are indisputable, it raises the importance of training such systems on unbiased data sets, x and y. In addition to RWD sets.

II. Data Science and Quality

There are issues with real world data (RWD) as often it is not generally collected for research purposes. Issues with the data includes non-rigorous data collection, non-purposive sample selection, episodic and / or incorrect timelines, containing data collection biases, reactive, and at best can only offer partial snapshots. As a result, RWD can also be generally messy and sparse, and requires statistically rigorous and valid methods to clean the data and employ error correct to overcome data inconsistencies. The process of careful data identification, prioritisation and inclusion, using both structured and unstructured data, can be critical for valid data analysis and subsequently real world evaluation (RWE). How issues of missing data are filled can often invalidate the findings and yet a system of independent data regulation and validation is not present. Thus the authors contend that transparency on the type and amount of error correction is essential to build more understanding and subsequently trust in AI systems. In the context of healthcare where often crucial information related to molecular biomarkers or end-points data can be missing, the missing data gaps may be filled by bridging to alternative data sources [48]. Data scientists are required to identify and adjust for confounding factors such as demographics, socioeconomics, psychographic and behavioural data. Further complications exist depending on the particular domain or even combination of domains in question. As such it is important for more explainable AI and research therein, to improve transparency of AI [41].

Genetic predispositions and / or Neurological processing may offer a baseline before conducting in-depth analyses. RWD is also subject to selection bias, as cohort selection and treatment decisions in clinical practice are not random. Essential to acquisition of relevant data assets, guidelines on design and validation of RWE studies can help in minimizing some sources of bias and inconsistencies [76]. In addition, standards for the development and maintenance of data assets needs to keep up with the rapid evolution of RWD. The use of legacy data systems may inhibit or prevent large scale predictive accuracy and yet be fundamentally important to the task at hand. The diversity and complexity of systems, data types, data locations and data availability means that there is often a lack of interoperability which heightens complexity for any data collaborations. At a micro level within organisations data is often held in different data systems. In such situations, there is undoubtedly a need to implement standardisation and maintain robust quality assurance (QA) quality control (QC) practices to support data robustness.

III. Current State of AI

It is generally regarded that deep learning, a subset of machine learning, is at the frontier of artificial intelligence research. Deep learning consists of multi-layered synthetic networks that are modelled on the human brain: namely, neural networks [68]. These deep learning networks have a nature of interoperability that is fundamental to the increasing potential of AI. Ideas about deep learning are not new, however. In 1943, [47] first discussed the notion of “neurons as elementary adaptive nonlinear processing units’’ [68] as opposed to logic-based units. Before this, analogue computers were thought of as logic-based, but with [47] ideas, the realm of possibility opened up. In the following years, computers had not quite reached the ability to analyse vast sets of data, though contemporary developments in technology have made this possible today.

Deep Face, for example, is a nine-layered neural network with over 110 million parameters created by Facebook that can identify human faces in photographs. Similarly, Deep Net which has been trained on over 150 million images from Google Net can identify facial similarities with accuracy levels of 98.73 and 96.12% [57]. However the issue of error correction amount and type is not presented or publicly available.

IV. Categorisation of AI

The assertion that Artificial intelligence (AI) seeks to process, understand and respond to data in the same ways which humans would [54] starts to become hyper complex considering the above human intelligence with the aim of AI to be anthropomorphic in nature and whereby the algorithms allow AI systems to mimic human cognitive functions to solve problems [38]. The next section considers some artificial intelligence fundamentals and as such the 4 basic AI concepts of **1,** Categorisation - where metrics are created relevant to the domain, then **2,** Classification - where the data is analysed to determine the most relevant to solving the problem, followed by **3,** Dimensions and types of intelligence, followed by **4,** Machine / Deep Learning which on a basic level involves anomaly detection, clustering, deep learning and analysis. Finally **5,** Collaborative filtering where patterns are detected across large data sets resulting in certain forecasts, predictions or entailments. A key consideration in any new classification of AI is that of interpretability and explainability, as is the key criteria suggested to establish cause and effect [41] in scientific theory and should be part of the AI decision making explanation [43] and advise "re-framing conversations around machine autonomy to foreground human actors and the broader sociotechnical context in which such systems are embedded''.

This following section seeks to break down commonalities among categorisation of AI; with the base level of Narrow / Weak AI - which lacks the ability to understand context but can perform simple demonstrative tasks;

At the central level is General AI; where it is able to understand context and make inferences from it and also operates on little to no information, and exhibit powers of reasoning and creativity; finally Super AI - which possesses an intellectual capacity far superior to that of a human beings. [36]. An additional consideration can be given to the methodology of AI whereby the training of it can be considered and includes; Training AI - can learn and improve over time; and Inference AI - requires human interference to make more relevant suggestions such that [55] "expertise in epistemology, critical thinking and reasoning are crucial to ensure human oversight of the artificial intelligent judgements and decisions that are made, because only competent human insight into AI-inference processes will ensure accountability". This method of classifying AI shares some similarities, but is advanced, with the following where AI has been classified is by the level of human systems interaction and includes; Supervised AI - which requires human monitoring and feedback; Unsupervised AI – the unsupervised suffers from the lack of "expert" touch (in the context of dermatology) during the training [1] and could be considered Black Box AI which does not require human interference, and Reinforcement AI whereby occasional human interference is needed. The combination method or the [1] "semi-supervised learning" method has also been introduced, which utilizes a small amount of labelled data and a larger amount of unlabelled data.

Beyond the above and to confuse matters further, there are also ways of classifying AI according to its potential:

- Expert systems, Analytical AI, Human-inspired AI, Humanised AI

The determining factors of which are listed as follows [42]

- Cognitive Intelligence
- Emotional Intelligence
- Social Intelligence
- Artistic Creativity

Clearly there are a number of similarities across the various labels and categories. In addition there are some basic yet important considerations pertaining to the level of human involvement, the type and level of intelligence that will feed forward into the design of the proposed model. However there are some gaps in relation to new areas of, in addition to new streams of intelligences, and their overall contribution to AI.

V. Dimensions of AI

In order to address issues of classifications there is a need to further review AI. Central to any discussion of AI should be human intelligence, Abstract Reasoning Corpus, Skill acquisition efficiency etc. However the current performance of AI should be considered as becoming far greater than human intelligence. A great deal is written about the abilities of AI systems to outperform humans in games, calculations and other narrow fields but, this is where key issues arise due to the differences in how humans might process the data in different domains. One of the key components of intelligence is cognitive learning which involves the acquisition of knowledge and internal mental structures through cognitive processes such as thinking, problem solving, language and information processing [69].

An interesting assertion [15] argues that intelligence cannot be measured by skill at a particular task or set of tasks. When we consider Abstraction and Reasoning Corpus (ARC) which is said to measure general fluid human intelligence, then a number of key elements need to be considered as important. These include that of skill acquisition efficiency which subsequently highlights concepts of scope, generalised difficulty, priors and experience [15]. However there are a number of omissions such as the beneficial process of trial and error, temporal and deep learning, and divergent vs convergent thinking in relation to the above. There is also the higher existential level of thinking. A further complexity arises with [25] theory of Multiple intelligences which implies the belief that there are to be a total of nine intelligences [53], where each person possesses a unique combination with one being the more primary or dominant variable The impact technology is increasingly having with people's ability to read [12] has an agonizing impact on [25] linguistic intelligence. Linguistic intelligence involves people with strong writing and speaking skills, memorization and reading [25]. Other types of the nine intelligences includes; Spatial, Bodily kinaesthetic, interpersonal, naturalist, music, linguistic, existential, logical-mathematical and Intrapersonal. Since fuzzy set theory was introduced by [82] in the 1960's, which suggested that uncertainty originating from human thinking can affect scientific problems. Since then, fuzzy logic has been successfully used in working with numerous practical applications. According to [65] "Fuzzy set theory is a research approach that can deal with problems relating to ambiguous, subjective and imprecise judgments, and it can quantify the linguistic facet of available data and preferences for individual or group decision-making'' [65].

Fuzzy set theory applied to psychology might be interpreted to suggest the cognitive processing is basically estimation rather than based upon thresholders, or reliable ground truth values. If enough people in a sample behave as if their strength of belief varies nearly continuously with the stimulus variable in the statement to be believed, then the given hypothesis would be supported and the psychological reality of fuzzy sets would be made more evident [34].

In the context of human intelligences, certain leaps across different logics and types of intelligences [25] can be intuitive, biased and not always with consistent or accurate entails resulting in sometimes best guess scenarios. However being able to draw upon multiple intelligences and logic processes in an intuitive way is advantageous in deducing entailments even with unfamiliar problems. This intuitive switch between the combination of previous rich experiences, different intelligences and logics results in human visualisations of entailments that often result in desired outcomes. This has a number of different underpinning axioms as a result of different methods for human learning. In addition to the above comes the further complexity that different logics (Proposition logic, First-order logic, Temporal logic, Probability theory, Fuzzy logic, etc.) or combinations of logic's that can be applied in order to determine what entails. Given the above overview of the complexity of human intelligence, it is highly unlikely that the capabilities of AI, is able to fully replicate the hyper complex human mind.

Deep learning, which is a process of AI, has made the most progress in solving complex problems consisting of recognizing speech from multiple speakers and identifying patterns in increasingly large data sets [64]. Subsequently, deep learning is already in some instances, substantially enhancing human capabilities [53]. Thus, with gaining knowledge instantly, this can have a profound effect on [25] multiple intelligence theory, where “logical-mathematic intelligence” is enhanced at a significantly faster pace than non-technological services. Thus, AI and other forms of technology are providing this instantly, and are of particularly benefit to scientists, mathematicians and philosophers who rely on this form of intelligence [25]. As such this context based problem solving could be considered as constituting narrow AI.

When intelligence is considered, it is important to discuss [21] second prominent theory of learning; behaviourism. Behaviourism can be defined as “changes in either the form or frequency of observable performance” [21]. The importance of behavioural learning implies the notion of learning from previous failures and the effects this has on the strengthening of future behaviour [73]. Conditioning occurs through interactions with the environment, thus behaviorists believe that the responses to environmental stimuli are accountable for future actions [26]. This importantly offers a perspective whereby the ground truth value may not act as a predetermined end point. The approach of behaviourism can be linked closely to reinforcement learning discussed later [75].

Consequently, the importance of learning through practice and encourages how it is imperative one must learn through the practice of skills before they can be performed accurately [39]. The usage of technology could be in fact diminishing bodily-kinetic intelligence. Face to face communication [17] can also diminish due to technology assisted interaction. This form of intelligence is prominent within athletes, dancers and surgeons, who are effective at body movement, hand-eye coordination and physical control [25]. Consequently, with technology constantly advancing, it provides the threat of the loss of “bodily-kinesthetic intelligence”. This discussion can be echoed with the theory of naturalist intelligence [25], the idea of nurturing and exploring with the environment. It can be feared that soon, with advancements of technology, this form of intelligence is unlikely to be diminished.

It could be considered that as our reliance on technology increased then certain types of human intelligence may diminish, whilst AI is exponentially increasing. One such area is that of behavioral learning, where intelligence can be heavily influenced by the notion of human interaction [77]. There has been particular discussion within literature on the negatives affects technology is creating affecting face-to-face communication. It has been found [8] that there had been a prominent decrease in face-to-face interaction amongst the youth, due to these individuals growing up with the internet as part of their everyday life during education, communication and entertainment. The lack of human interaction and communication would subsequently affect [25] “interpersonal intelligence’ dimension. Interpersonal intelligence can be defined as people with good interactions and communication skills, thus with a strong understanding of people and the emotions and motivations surrounding them. People with a lack of verbal communications and interactions face a significant threat of a core loss of intelligence.

A potential solution to the intelligence issue is through the use of cognitive architectures. Instead of aiming to create systems that are skilled in one aspect of human cognition in limited contexts, architectural research can be used to provide rich guidance across multiple tasks and domains [44]. Cognitive architectures differ from expert systems as they provide “counts of intelligent behavior at the systems level, rather than at the level of component methods designed for specialized tasks’’ [44].

The idea of cognitive architectures is not new. Most prolifically, [51] has argued for their existence in his “twenty questions paper’’ with his program for cognitive modelling.

Conversely deep learning and voice recognition, response and translate systems, within the area of verbal communications and interactions, AI powered systems may prove beneficial for both life and business improving impact of voice assistants such as Apple’s Siri, Amazon’s Alexa and Google’s Assistant are intertwined with the notion of deep learning networks, recognizing requests and providing instant answers [32]. Voice assistants are constantly being used to bridge the information gap between the ability to read and type, thus benefiting dementia sufferers providing a present voice willing to answer questions repeatedly without losing patience [32]. Additionally, voice assistants are constantly enhancing translation [32]. Google has recently launched a new set of earbuds, providing people with voice assistance for real-time voice translation, allowing users to gain hands-free audio translations [32]. Thus, it can be highly discussed the enhancement technology creates with human capabilities through the use of knowledge, thus emulating forms of intelligence [64].

VI. Issues with Baseline / Ground Truth Value

The importance of baseline conditions in scientific research is a prominent part of literature on research techniques. A baseline serves as an important reference point from which progress can be tracked and is an integral part of

Monitoring and Evaluation (M & E) frameworks [50]. It can be referred to at various points of the study process, from preliminary hypothesis testing to mid-point reviews. A baseline condition allows for comparisons to be made [29]. There are also number of stakeholders, including governments, citizens, the private sector, Non-Ggovernment Organizations (NGOs), Civil Society, international organizations, among others, which are now focused on increased performance of policies, programmes, and projects, which calls for enhancing baseline ground truth accuracy via results-Based Monitoring and Evaluation (RBME). Fundamental assumptions in experimental research are (a) the components and parameters of the conditions are known, and (b) those conditions are implemented with consistency and accuracy [3]. These assumptions apply to baseline conditions as well as intervention conditions. Within experimental research, it is widely regarded that baseline conditions are important for determining end points. A baseline serves as a reference point and is dependent on the following conditions: that "(a) the components and parameters of the conditions are known, and (b) those conditions are implemented with consistency and accuracy [31]. Within AI research, there are no consistent parameters. Studies in the field of AI (list some) test the power of AI systems not by their ability to be AI, but to achieve pre designed goals.

It is stated [7] there is an issue which is commonly faced when dealing with Web-based concepts: namely, vague ontologies. Within the realm of computer science, ontologies are defined as "the definition of domain concepts (extensions) and the relations between them'' [7]. To say that an ontology is vague is when concepts are not clearly defined, i.e with vague language. For example, in the case of this paper, Basic, Moderate, Complex and Advanced AI could be described as vague ontologies, since they do not reveal the intricacies of component parts. Frege (1906) cited it [60] highlights that a "concept must have a sharp boundary'' in order to avoid doubt [60].

A key question pertains to how error is minimized in machine learning algorithm. This is being done through the running of the algorithm as many times as is necessary in order to compare model prediction with the ground truth value with subsequent adjustments of parameters that result in smaller error, over time the result should be an increase in prediction accuracy. However, when there are a multitude of possibilities for the initial error, with varying degrees of impact, then simply adjusting parameters is unlikely to be sufficient. In the context of reinforcement learning with neural networks whereby the networks based agent discovers complete quantum-error-correction strategies, there is still a need for measurement outcomes [24]. They also suggest that the combinatorially large search space presents a substantial challenge when attempting to find Q-E-C strategies from scratch without human guidance [24]. The suggested solution is a tier stage learning reliant on human guidance [24]. The optimization of algorithms which characterises machine learning , through guessing and guessing again until close to ground truth value is insufficient in comparison to deep learning.

The programming towards a predictable ground truth value negates the possibility of an alternative end point. The reliance on software code and the numerous parameter adjustments makes the processing and final outcome relatively distorted particularly when applied within unacceptable error domains. This pertains to non-acceptable error margins between the prediction and the ground truth values which on a mass scale may mean a small percentage of unacceptable errors. In addition whilst there are a number of halo claims in relation to AI, in the main there are still substantial hurdles for AI to overcome before it is to be wholly relied upon. The premise here is that the capability of AI does not sufficiently compare to the breadth and hyper complexity of human thinking but only mirrors and advances narrow types of human thinking.

Deep Learning could be considered as better than Machine Learning due to the needlessness of Feature Extraction. Machine Learning models use feature extraction to determine whether a given picture shows a car or not and must first have the features of a car (shape, size, windows, wheels etc.) must be programmed into the algorithm. It was found that having latent features extracted using DSAE proved useful for driving behaviour visualization [6]. However if it important here to highlight the previously mentioned issues with data. A key to Deep Learning models is that they increase their accuracy with the increasing amount of training data, whereas traditional machine learning models such as SVM and Naive Bayes classifier stop improving after a saturation point. In relation to algorithm consistency and convergence then it becomes important that a "Bayes network belief propagation algorithm converges on a probability distribution dictated by probability theory or proving that a theorem prover is sound and complete with respect to a semantics for some logic" [13].

In addition, recognition of the problems associated with neural networks being incredibly large and built by hand, has now been replaced with a technology which combines both human and AI to build the neural networks. As such this combination now accelerates the deep learning design for a number of applications including autonomous driving (Article – Researchers find a way to Harness AI Creativity-Dramatic performance boost to Deep Learning). By leveraging human ingenuity and experience with the meticulousness and speed of AI has a major contribution to on-the-edge deep learning solutions (Article – Researchers find a way to Harness AI Creativity-Dramatic performance boost to Deep Learning).

The application of deep learning methods have resulted in impressive advances in NLP, especially in the development of unsupervised models using recurrent neural networks and auto encoders that reduce dependence on high-quality, manual annotations of text data [18]. The methods of applying deep learning on electronic health records in Swedish to predict healthcare-associated infections [35] allows algorithms to learn high-level abstractions from clinical data and notes when concepts are not mentioned explicitly [66]. Availability of large volumes of real-world clinical data enables the training, development, and validation of new algorithms [71].

Beyond NLP, advances in machine learning have enabled new approaches for prediction of disease onset and future diseases [49]. This is in addition to the [81] exploration of machine learning techniques in predicting multiple sclerosis disease course. Another application is in image recognition for classification of radiology and pathology images [58]. Further applications include [28] development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs [22]. Dermatologist-level classification of skin cancer with deep neural networks, methods for assessing disease heterogeneity and predicting patient outcomes, given the information about a patient, their history, and individual-specific variability, have demonstrated capabilities to include both observed as well as latent features extracted from messy, multivariate EHR data [63]. Advanced analytics using machine learning on longitudinal RWD has the potential to inform and reframe drug development and clinical trial design strategy—through patient stratification into subgroups based on disease subtypes, drug treatment efficacy, progress, side effects, and toxicity profiles—by shifting from presumption of a single disease to multiple, related diseases. As machine-learning algorithms and frameworks continuously advance, there will be improvements in the ability of these models to learn continuously as new information emerges either in the form of additional data sources or updated treatment guidelines [19].

However there are issues in the application of NLP according to [71] where NLP methods is one approach to enable extraction and conversion of unstructured information from clinical text data to structured observations, to extraction of findings within ejection fraction from laboratory reports, biomarker information from pathology reports. In addition, the capture and use of patient characteristics such as emotional and social behaviors from physician notes [71]. Furthermore there is a central problem in that predefined fields in EHR (e.g., problem lists, past history, or test result fields) capture only certain disease information and may miss the trends of other prevalent, but unlisted, health conditions. NLP can be a powerful tool to extract symptoms from physician notes or textual data from lab reports to enable identification of those trends/conditions, thus complementing the assessments using structured data. The thoroughness and cross examination capabilities of a fully trained physician offers a multi-level health examination with a safety net consciousness such as AI safety [78] whereby a possible solution is where necessary verification with physicians can be sought. Thus beyond current AI capabilities.

The intelligence puzzle (as coined by [13] is an obstacle unique to the development of AI systems that seek to replicate "human-level artificial intelligence" (HLAI). In current science, we cannot say how "a system composed of unintelligent parts (such as neurons or transistors) can behave as intelligently as people'' [13]. Thus it may be difficult to project this understanding of natural laws unto artificial laws and seek to recreate these systems in technology.

Furthermore, human intelligence is implicated in several fields: from economic policy to organisational practice and culture and society. As such, HLAI should be considered an atypical science [13]. Reinforcing this notion is that the goals of HLAI are more ambitious than the goals of its scientific counterparts. For example, medicine is often described as "the science or practice of the diagnosis, treatment, and prevention of disease" as opposed to "artificial immortality'' (p.38). Resultantly, progress in achieving the latter seems minimal compared to the former: if we were to give AI (or HLAI in this case) a similarly defined goal, there would be fewer questions of progress made in the field [13].

It is not uncommon for terms that have entered the public domain from academia to be followed by clouds of misunderstanding and confusion [40]. In the case of HLAI (or AI in general), this bewilderment extends to the scientists involved in the field. Simply put: "the idea that our era is somehow seeing the emergence of intelligence in silicon that rivals our own entertains all of us, enthralling us and frightening us in equal measure. And, unfortunately, it distracts us'' [40].

Scholars [80] argue that NLP is no longer valid or relevant, however. This is for the following reasons: 1. It is formed based on speculation "that the mind processes information at or below the Turing Limit''. That it overlooks the (purported) reality that cognition surpasses computation. In its place, there should forms of "cognitive modelling untrammelled by standard computation'' (p. 627) such as elements of hypercomputation like "analog chaotic neural nets, trial-and-error machines, Zeus machines'' and so on. Supporting this notion are [10] and [67].

A slightly less critical view of NP comes from [74] who argue that while Newell's ideas about an amalgamated Theory of Mind (ToM) are not wrong, his methodology is flawed. Rather than altering existing approaches to meet criterion within NP, we should hold the complexities of the human mind in high regard, while developing "complementary theories at both psychological and connectionist levels, and cross-validate them'' [74].

VII. Cross Domain Intelligence

An important area for consideration regarding a new approach to AI classification has to come from cross domain thinking. This includes the use and modelling of intelligence via examples from the natural world. There are a number of examples where significant breakthroughs regarding thinking and problem solving has been better achieved from things that exist in the natural world. For example there have been profound breakthroughs using Lobster (meridional) Eye Technology; for nanotechnology [56], Materials Science and in particular laser technology [46], Scientific Apparatus such as telescopes [16], [33], Physics [4]. In relation to the development of algorithms for predictive modelling, based on the 'collective intelligence' concept, contributions have come from Entomology and Ant Colony Optimisation [37] and Particle Swarm Optimisation, are the two most commonly known nature inspired algorithms. In addition the use of Bee's for Artificial **bee** colony-based **predictive** control for non-linear systems [62]. An additional contribution comes from the context bio-inspired computation algorithm, in particular AI based optimization algorithm inspired from the nature of vortex [52]. It is the authors' belief that new classifications should cater for the various realms and combinations thereof, that has potential for beyond human intelligences.

VIII. Key Themes on AI and Intelligence

There are a variety of issues pertaining to AI that includes not having any agreed standard categories from which to bind the different levels of AI. An additional set of issues is the lack of training and subsequent knowledge within the domain of human intelligence. Furthermore, there are issues with Data accuracy, Data issues with combinationally large search space. There are issues with quantum-error-correction strategies in addition to prediction errors against ground truth values in a variety of domains as presented herein. In particular the transparency of the level and type of error correction strategies should be critically considered particularly in critical domains such as healthcare. Early warning scoring [30] might enable contingencies to be enacted. The future direction and potential from cross domains has been discussed and remains a key area to consider in the AI classifications. In addition the authors recognise the level of human interaction in the design and management is also important. The above categorisations and definitions of AI offers a basis for informing a better standardisation of AI, and help to fulfil the purpose of this paper. Given the variability of the above the authors propose that there will be some level of confusion and lack of understanding from the data scientists' panel.

IX. Research Panel and Sample

Since this paper seeks to represent perceptions of AI, from the experts in data science, a non-random purposive sample was used. Data scientists/ experts were contacted to ensure theoretical, logical and analytical assumptions could be made, by applying expert knowledge to a cross-section of the population. Purposive sampling was used in order to ascertain perceptions from those working in the field of data science. The domain of data science being a thorough industry with its own unique screening process. By asking data scientists/ experts, the authors hope to gain an insight into how industry perceives, approaches, and classifies AI.

The authors have defined data scientists/ experts as those who are employed by corporations to work in machine learning, software development and traditional research, with responsibilities that can include data mining, algorithm development and/or data managing. By capturing the opinions of data scientists, modern perceptions of AI in the workplace can be gathered effectively. This is because, due to the comprehensive nature of data science, it is possible to gain insights into computer science, business knowledge and statistics respectively.

X. Research Findings

We surveyed a group of experts (both male and female) aged 23-45, in professions such as data mapping, software engineering and data science. The majority of participants have been directly involved with AI for at least 2 years, with 4 being involved for 3+ years. When asked, "To what extent are you involved with AI?" on a scale of 1 (low involvement) to 5 (high involvement), participants had, on average, reported 3 (moderate involvement). When asked, "Do you consider yourself as having an advanced understanding of AI?", only 5 out of 20 answered with a "yes". Though this tended to correspond with the level of involvement, this was not always the case. For example: one participant who had been working with AI for 3+ years (with a high involvement of 4) answered this question with "to some degree", while another with high involvement (4) answered with "no'' when asked about their understanding of AI.

Therefore, the authors conclude that length of time and extent of involvement do not necessarily relate with ideas about own understanding. While this uncertainty could be due to modesty or the general notion that "true knowledge is knowing you still have more to learn'', it points to the wider issue of ambiguity behind the umbrella term "artificial intelligence''. This supports the general claims made in this paper so far, and calls for a revision of terms. As a general observation, responses tended to become less specific throughout the various stages of the questionnaire, more specifically, as terms progressed through basic, moderate, advanced and complex AI.

Question 1: What do you consider as low level basic AI?

When asked "what do you consider as low level basic AI?'' responses ranged from "smartphones'' to "supervised learning within a single-layer neural network''. These variances indicate that even at a basic level, there is a lack of general consensus as to what constitutes AI. Only the minority of participants answered "Machine learning", "Automated learning", "Supervised learning via a neural network", akin to a combination of basics of AI and complex AI. Another example provided was "AI based on inputs and outputs'' and "building a playable video game containing an AI opponent (The complexity of the game will correspond to the "level' of the AI"). The former suggests an understanding of grounding principles of machine learning while the latter indicates an idea of creativity being an important parameter. It also indicates that the individuals surveyed have low confidence and a low understanding of what constitutes AI. Another minority answered with Consumer products and services, such as Netflix, Facebook Messenger and Spotify Discover Weekly were all listed as being low level basic AI, as well as product features such as "red-eye reduction'' and "camera lens focus''. This consumer product answers were all given as a stand-alone halo answer without any real underpinning justifications. In addition to the above, "automation of tasks'' was mentioned as low level basic AI, as well as "AI based on inputs and outputs''. This notion ties into ideas about basic AI being operation-led (list scholars here) and designed with a clear goal in mind. The limitation can be extrapolated to be the inability to deviate from the goal, or to think independently. Hence, the element of independent thought will be carried forward to inform the overall design and level of AI.

More statistical functions such as "clustering'' were also given, as well as "social media algorithms'' and search/ return functions like "metadata tagging''. This ties into the literature on AI that notes how "in their simplest form, intelligent agents are merely programs that solve specific problems''. Already, it is visible that perceptions of what constitutes low level basic AI are, superficial in some cases, inaccurate and overall divided.

Question 2: What is moderate level AI?

When asked about moderate level AI, learning emerged as a prominent theme in the form of "trained systems'', "learning algorithms'' and "supervised and unsupervised learning''. In some cases, examples were given, for example, "regression, classification and decision trees'' (within supervised and unsupervised learning) and "multi-layer neural networks'' (within unsupervised learning specifically). Similar to the aforementioned, "structural equations'' were also listed as being an example of moderate level AI, introducing modelling as a concept.

In one case, the concept of forecasting was noted, with the answer of "using AI to predict the stock market''. This indicates that using imperfect (incomplete) information is a factor in an AI system's complexity, and is an element which will be carried forward to later sections. Similarly, another respondent answered with "boosted methods, ensemble methods'' (meta-algorithms), indicating that bias reduction and combined ML techniques differentiate basic to moderate AI, additional factors to consider when making recommendations.

Specific examples of consumer products appeared throughout the answers, such as Google Home, Amazon Alexa and IBM's Watson. Similar to branded consumer products, functionalities such as "voice search'' were also listed, as well as unbranded goods, like "smart homes''. This follows on from the findings in Question 1, since it reinforces this notion that individuals tend to think of AI in terms of its benefits and personal gain. In one case, AlphaGo was listed as being an example of moderate AI (thus indicating a theoretical as well as commercial understanding of AI systems), though the former greatly surpassed the latter in volume. One participant even answered with "not sure as I'm unfamiliar with the different levels of AI'', suggesting a need for universal definitions.

Question 3: What do you consider as complex AI?

When asked "What do you consider as complex AI?'', three respondents listed "neural networks'' as an example. In one respondent's answer, this was accompanied with a use case, specifically: "utilising unsupervised neural networks to provide simple value, i.e. recommendations''. In another, "deep neural networks / unsupervised learning'' were mentioned together. Unlike previous answers, "deep'' appeared as a distinctive factor (at least in the respondent's mind) of what differentiated moderate AI and complex AI.

Once again, learning emerged as a prominent theme, but with a new dimension of "unassisted learning/ direction''. This ties into the aforementioned operation on imperfect information, and will be considered in later sections. In a similar manner, one respondent listed "AI based on Artificial learnings and little to none input'' as an example of complex AI. This supports classifications from (insert scholar from above). From this, themes of independent thought can be extrapolated, as a component of complex AI.

Following on from independent thought, an additional element of creativity can be extrapolated, as seen in the case of "coding/ creating/ consulting'' and "creative machines''. This supports ideas about creativity being a feature in the classifications of more developed systems (Kaplan and Haenlein 2019). Beyond creativity, "voice emotion identification'' was also introduced as a concept as well as "image'' and "speech'' recognition. This indicates that perceptions of complex AI systems entail the processing of unstructured data.

Question 4: What do you consider as advanced AI?

When asked about advanced AI, respondents tended to become less confident in their answers. For example, one participant responded with "robots?" and another with "something beyond that of a human mind". Several times, "I don't know'' or "same as before'' was listed as the answer, indicating uncertainty in own understanding, while the answers "creating AI'' and "a fully conscious being - not there yet'' indicate doubt in the field as a whole in terms of progress. However, unlike previous questions, one participant did answer with "General Intelligence", indicating an understanding of existing AI classifications. Other respondents did not refer to such classifications, indicating a lack of general consensus, even among experts in the field.

Once again, independent thought and creativity were raised as important attributes, with "AI based solely on its own learnings without input from humans'' and "creative systems able to think independently''. For another respondent, the similarity of AI systems compared to a human baseline was an indicator of an AI systems advancedness: "creating a machine to have a similar level of rounded intelligence like a human. Enabling it to think logically, learn and grow''. Consciousness was also introduced as a concept, as well as "full autonomy'': introducing the notion of active brain state achieved and self-directed goal pursuits. This human tendency to anthropomorphize could serve as an important measure of identifying (universally) what constitutes AI systems of varying degrees of complexities, thus will be carried forward.

Overall: Questions Summary

From doing the above analyses, it is apparent that there is little to no shared consensus of what constitutes varying degrees of AI classifications. Albeit, there are clear themes within each section, indicating some transference of thinking, though these are not guaranteed. This could be due to vague ontologies in the questionnaire design such as "low level" or "advanced'', words which are subjective in nature - an error which the authors acknowledge. However, in search for a better word - this reinforces the need for a collective set of definitive parameters which can be used by individuals to form their ideas upon. There are potentially lexical gaps that need filling, or clear boundaries set from a respected source that help guide and inform the answers of individuals.

XI. Conclusions

Conclusion – do the parameter adjustments need to be transparent re AI?

Conclusion – better training of data scientists to eliminate ground truth value differences and of professionals to be aware of the compounded adjustments is a recommendation of this paper. When considering the vast pace, various labels and variety of AI, combined with the complexity of the domain of human intelligence, the responses by the data scientists becomes partially explainable. In addition the domain of data science with the roles within it being generally narrow, offers another explanation of the paucity of response.

In addition the low confidence displayed by the data scientists should have been expected given the lack of training on the domain of human intelligence. The traditional background of data scientists is generally that of computer science. This is also in relation to the rate of change.

The current classification and diversity of terms and labels surrounding AI are too broad and overarching and may partially explain the lack of clarity and ease of which the public the attaches their worst case scenario's to the concept. The fog which surrounds AI is due in part to the hyper complexity that exists and the range of labels, categories and classifications that reduces transparency both in the research being conducted and the how it is communicated. Furthermore the use of a variety of realms from which to draw new, beyond human intelligences adds a further complication and challenge that needs to be addressed.

The concept of a sharp boundary coined by Frege [60] in 1906 applied to AI classification is further increased by the following, Hierarchy of AI. This the authors believe better captures the variety, complexity, interoperability and future of AI.

Subsequent use of the hierarchy of AI and the associated labels will provide greater transparency leading to improved understanding and the locating of research (both national and international) being undertaken within it.

XII. Dimensions of AI

The following model (Fig. 1) demonstrates broadly the proportionality and scope of intelligences at each level. Overall it provides relatively sharp boundaries [60] between the levels. It caters for the latest developments and interoperability of AI in addition to the future potential. Thus acts as a basis for identifying potentially confusing areas and in some cases dangerous developments within AI, thus serving an agenda for understanding, regulation and transparency. It is important to note that the authors depict the hierarchy in the shape of an inverted pyramid to better capture the nature of AI. The following provides a level of explainability regarding the Hierarchy of AI. The first level is termed Systems intelligence (Table I shows the factors that constitute this level). The second level is termed Neural intelligence (Table I shows the factors that constitute this level). The top level considers the nature of multi-layered synthetic networks and interoperability thereof, whereby the systems and algorithms have the capability to exchange and make use of a variety of information types, across new boundaries, extends the level of risk, including risk censoring [20], and difficulty regarding the potential transparency of the error correction strategies. That is in addition to potential outcomes. As such the authors highlight this type of AI at the top of the AI Hierarchy. This type is termed Transversal intelligence. The hierarchy importantly depicts the level and combination of multiple error correction strategies due to the interoperability across data systems and data types. As such this Transversal level of intelligence is where it becomes imperative that there is a high level of regulation and transparency.

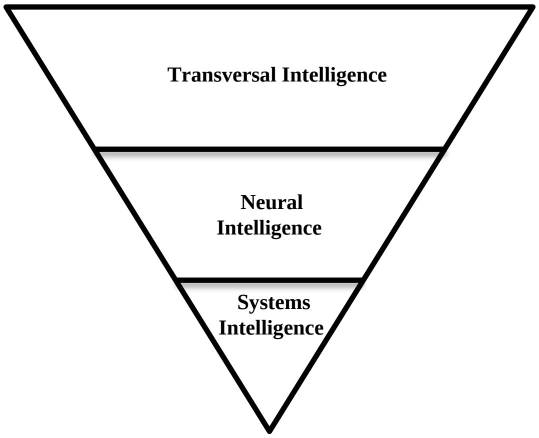


Fig. 1. Model: Hierarchy of AI.

TABLE I. Explanations of the levels and criteria within the Hierarchy of AI

Hierarchy of AI		
Level	***Overview***	**Key Parameters**
Transversal Intelligence	Creative Intelligence	Typically no – minor human involvement. Range of ability to intelligently expand the data sets to new domains automatically. Extreme level of interoperability. Ability to determine new; end points, rules, laws and beyond human intelligence.
Neural Intelligence	Adaptive Intelligence	Typically no - Moderate human involvement. Range of ability to intelligently expand the data sets within human set parameters and across human domains. Machine and Deep Learning interoperability across human levels of data abstraction. Enabling Forecasting and Predicting within the scope of human intelligence. Combines Narrow, deep and broad domains informing results.
Systems Intelligence	Non adaptive Systems Intelligence	Typically hi level of human involvement. In ability of systems to expand the data sets automatically. Narrow and shallow in domains.

References

[1] Acharya, P. and Mathur, M. (2020), Artificial intelligence in dermatology: the "unsupervised" learning. Br J Dermatol. Accepted Author Manuscript. doi:10.1111/bjd.18933

[2] Agre, P. (1997) "Toward a critical technical practice: lessons learned in trying to reform AI". In Beyond the Great Divide: social science, technical systems, and cooperative work, Edited by: Bowker, G. C., Gasser, L., Star, S. L. and Turner, W. 131–58. Mahwah, NJ: Erlbaum.

[3] Anderson, J.R., and Lebiere, C. (2003) The Newell Test for a Theory of Cognition, Behavioral and Brain Sciences, Vol.26, pp587-640

[4] Aslanyan V; Keresztes K; Feldman C; Pearson JF; Willingale R; Martindale A; Sembay S; Osborne JP; Sachdev SS; Bicknell CL; Houghton PR; Crawford T; Chornay D, The Review Of Scientific Instruments [Rev Sci Instrum], ISSN: 1089-7623, 2019 Dec 01; Vol. 90 (12), pp. 124502; Publisher: American Institute Of Physics; PMID: 31893794

[5] Bart A., and Feigenbaum EA,. (2014), The handbook of artificial intelligence,

[6] Bichicchi, A. et al. (2020) 'Analysis of Road-User Interaction by Extraction of Driver Behavior Features Using Deep Learning', IEEE Access, Access, IEEE, 8, pp. 19638–19645. doi: 10.1109/ACCESS.2020.2965940.

[7] Bourahla M. (2015) Exact Reasoning over Imprecise Ontologies. In: Amine A., Bellatreche L., Elberrichi Z., Neuhold E., Wrembel R. (eds) Computer Science and Its Applications. CIIA 2015. IFIP Advances in Information and Communication Technology, vol 456. Springer, Cham

[8] Brignall and Van Valley (2005) the impact of internet communications on social interaction, Volume 25, 2005 - Issue 3,

[9] Bringsjord,S. and Schimanski B,. (2003) What is Artificial Intelligence, psychometric AI as an answer, Proc. of the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003)

[10] Bringsjord, S. and Zenzen, M. (2003) Superminds: People harness hypercomputation, and more, Studies in Cognitive Systems, Springer, Vol.29,

[11] Caliskan, C, Bryson, J.j. Narayanan, A (2017) Semantics derived automatically from language corpora contain human-like biases, Science 14 Apr, Vol. 356, Issue 6334, pp. 183-186

[12] Carr, N., 2008. Is Google making us stupid? What the Internet is doing to our brains (Vol. 1). July

[13] Cassiamatis, N.L. (2012) "Artificial Intelligence and Cognitive Modelling Have the Same Problem", Theoretical Foundations of Artificial General Intelligence Atlantis Thinking Machines, vol. 4, pp. 11-24,

[14] Cave, S., Dihal, K. (2019) Hopes and fears for intelligent machines in fiction and reality. Nat Mach Intell 1, 74–78

[15] Chollet, F. (2019) On the Measure of Intelligence, arXiv:1911.01547 [cs.AI]

[16] Daniel, V. et al (2019). In-Orbit Commissioning of Czech Nanosatellite VZLUSAT-1 for the QB50 Mission with a Demonstrator of a Miniaturised Lobster-Eye X-Ray Telescope and Radiation Shielding Composite Materials, Space Science Reviews. Aug 2019, Vol. 215 Issue 5, pN.PAG-N.PAG. 1p. DOI: 10.1007/s11214-019-0589-7,

[17] Drago, E., (2015). The effect of technology on face-to-face communication. Elon Journal of Undergraduate Research in Communications, 6 (1).

[18] Dubois, S., Romano, N. (2017) Learning effective embeddings from medical notes, arXiv preprint arXiv:1705.07025.

[19] Dorajoo, S.R. & Chan, A. (2018) Implementing clinical prediction models: pushing the needle towards precision pharmacotherapy. Clin. Pharmacol. Ther. 103, 180– 183

[20] Dyagilev K, Saria S. (2015) Learning (predictive) risk scores in the presence of censoring due to interventions. Mach Learn 2016;102:323–48.doi:10.1007/s10994-015-5527-7

[21] Ertmer, P.A. and Newby, T.J., (1993). Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective. Performance improvement quarterly, 6(4), pp.50-72.

[22] Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639):115–118.

[23] Fogel D.B. (1995) Evolutionary Computation Toward a New Philosophy of Machine Intelligence, 3rd Ed, Wiley and Sons

[24] Fosel, T., Tighineanu, P,. Weiss, T., and Marquardt, F. (2018), Reinforcement Learning with Neural Networks for Quantum Feedback, Physical Review X, Vol.8, Issue 3, July-Sept.

[25] Gardner, H., 1992. Multiple intelligences (Vol. 5, p. 56). Minnesota Center for Arts Education.

[26] Greer, D. R., Dudek-Singer, J. and Gautreaux, G. (2006), Observational learning. International Journal of Psychology, 41: 486-499. doi:10.1080/00207590500492435

[27] Gudwin, R.R. (2000) Evaluating intelligence: a computational semiotics perspective, Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics. 'cybernetics evolving to systems, humans, organizations, and their complex interactions'8-11 Oct 2000.

[28] Gulshan V, Peng L, Coram M, et al. (2016) Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016;316(22):2402–2410. doi:10.1001/jama.2016.17216

[29] Gusnard, D.A. and Raichle, M.E. (2001) Searching for a baseline: functional imaging and the resting human brain, Nature Reviews Neuroscience, Vol.10, pp685-694

[30] Henry, K. E, Hager, D. N, Pronovost, P. J, and Suchi, S. (2015) A targeted real-time early warning score (trewscore) for septic shock. Science Translational Medicine, 7(299 299ra122): 1–9,

[31] Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The Use of Single-Subject Research to Identify Evidence-Based Practice in Special Education. Exceptional Children, 71(2), 165–179.

[32] Hoy, M.B., 2018. Alexa, siri, cortana, and more: An introduction to voice assistants. Medical reference services quarterly, 37(1), pp.81-88.

[33] Inneman, et al, (2019) see Nentvich, O. j., & Urban, M., & Blažek, M., & Inneman, A., & Hudec, R., & Sieger, L. (2019). Lobster eye optics: position determination based on 1D optics with simple code mask. 31. 10.1117/12.2528505.

[34] Jackson, S., and Jacksina, C.M. (2018) Impact of Fuzzy Techniques in Psychology, International Journal of Recent Research Aspects, April 2018, pp491-494

[35] Jacobson, O., and Dalianis, H. (2016) "Applying deep learning on electronic health records in Swedish to predict healthcare-associated infections", Proc. 15th Workshop Biomed. Natural Lang. Process., pp. 191-195,

[36] Jajal, T. D. (2018) Distinguishing between Narrow Ai, General Ai, and Super Ai, Medium, Artificial Intelligence. Accessed online https://medium.com/@tjajal/distinguishing-between-narrow-ai-general-ai-and-super-ai-a4bc44172e22

[37] James, L. (2018) 5 Ways mother nature inspires artificial intelligence, Towards Data Science, Jan 14, 2018. Accessed 12/02/2020 https://towardsdatascience.com/5-ways-mother-nature-inspires-artificial-intelligence-2c6700bb56b6

[38] Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H. and Wang, Y., (2017) Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology [online], 2 (4), 230-243.

[39] Johnson, R.T. and Johnson, D.W., 2008. Active learning: Cooperation in the classroom. The annual report of educational psychology in Japan, 47, pp.29-30.

[40] Jordan SC, et al. (2019) Daratumumab for Treatment of Antibody-Mediated Rejection in a Kidney Transplant Recipient [abstract]. Am J Transplant. 2019; 19 (suppl 3).

[41] Kade, L., and Maltzan, S.V. (2019) Towards a Demystfication of the Black Box – Explainable Ai and the Legal Ramifications, Journal of Internet Law, September

[42] Kaplan A, Haenlein M. (2019) Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons; 62:15.

[43] Krafft PM, Young M, Katell M, et al. (2019) Policy versus Practice: Conceptions of Artificial Intelligence. SSRN Electronic Journal. DOI: 10.2139/ssrn.3431304.

[44] Langley, P. (2019) Explainable, normative, and justified agency. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, 2019.

[45] Legg, S. and Hutter, M. (2007) Universal Intelligence: A Definition of Machine Intelligence, Minds and Machines, Vol.17. Issue 4, pp391-444

[46] Lin, Kaijie; Yuan, Luhao; Gu, Dongdong. (2019) In Journal of Materials Processing Tech.. May 2019 267:34-43 Language: English. DOI: 10.1016/j.jmatprotec.2018.12.004, Database: ScienceDirect

[47] McCulloch, W.S. and Pitts, W. (1943) A logical calculus of the ideas immanent in nervous activity, The Bulletin of Mathmatical Biophysics, Vol.5, Issue 4, pp115-133

[48] Miksad, R.A. & Abernethy, A.P. (2018) Harnessing the power of real-world evidence (RWE): a checklist to ensure regulatory-grade data quality. Clin. Pharmacol. Ther. 103, pp.202–205.

[49] Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T. (2016) Deep learning for healthcare: review, opportunities and challenges, Briefings in Bioinformatics, Volume 19, Issue 6, November 2018, Pages 1236–1246, https://doi.org/10.1093/bib/bbx044

[50] Moses, O. S. and Ssekamatte, D. (2016) Using baseline studies as a basis for monitoring and evaluation: A review of the literature, The Ugandan Journal of Management and Public Policy Studies Volume 11 Number 1 November 2016.

[51] Newell, A. (1973) You can't play 20 questions with Nature and win: projective comments on the papers of this Symposium, May 1973

[52] Onet, E.V. and Vladu, E., 2008. Nature inspired algorithms and Artificial Intelligence. Journal of Computer Science and Control Systems, (1), p.66.

[53] O'Neill, C. (2017) "The Ivory Tower Can't Keep Ignoring Tech." The New York Times, 17-Nov

[54] Oswald, E., 2019. What is artificial intelligence? Here's everything you need to know [online]. Digital Trends. Available from: https://www.digitaltrends.com/cool-tech/what-is-artificial-intelligence-ai/ (Accepted 9 March 2019)

[55] Pedersen, T. & Johansen, C.,(2019) Behavioural Artificial Intelligence: An agenda for systematic empirical studies of Artificial Inference, Ai and Society, Springer London, pp1-14 https://doi.org/10.1007/s00146-019-00928-5

[56] Peng, J.S., and Cheng, Q.F. (2017) High-Performance Nanocomposites Inspired by Nature, Advanced Materials, Vol.29, pp1-16

[57] Prasad, P., Pathak, R., Gunjan, V., & Ramana Rao, H.V. (2019). Deep Learning Based Representation for Face Recognition. 10.1007/978-981-13-8715-9_50.

[58] Ramagopalan, S.V et al (2017) Risk of Thrombosis in Sjögren Syndrome: The Open Question of Endothelial Function Immune-mediated Dysregulation, Luca Quartuccio, The Journal of Rheumatology Aug 2017, 44 (8) 1106-1108; DOI: 10.3899/jrheum.170462

[59] Razavian, A.S., Sullivan, J., Maki, A., and Carlsson.S. (2016) Visual instance retrieval with deep convolutional networks. CoRR, abs/1412.6574,

[60] Ricketts, T. (1997) Frege's 1906 Foray into Metalogic, Philosphical Topics, Vol.25. Issue 2, Fall, pp169-188

[61] Russell S.J. and Norvig P. (2016) Artificial Intelligence : A Modern Approach, Pearson Education Limited, London

[62] Sahed, O. A., Kamel, K., and, Abousoufyane. B. (2015) Transactions of the Institute of Measurement & Control. Jul2015, Vol. 37 Issue 6, p780-792. 13p. DOI: 10.1177/0142331214546796

[63] Schulam, P., Wigley, F., and Saria, Suchi, S. (2015). Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence,

[64] Sejnowski (2019) http://theconversation.com/artificial-intelligence-will-make-you-smarter-101296

[65] Shan, M., Chan, A. P., Le, Y., & Hu, Y. (2015). Investigating the effectiveness of response strategies forvulnerabilities to corruption in the Chinese public construction sector. Science and EngineeringEthics, 21(3), 683–705.

[66] Shickel, B., Tighe, P. J., Bihorac, A. & Rashidi, P. (2017) Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J. Biomed. Health Inform. 22, 1589–1604.

[67] Siegelmann, H.T. (1999) Neural Networks and Analog Computation. Beyond the Turing Limit, Springer, Switzerland.

[68] Sinz, F., & Bethge, M. (2008). The conjoint effect of divisive normalization and orientation selectivity on redundancy reduction in natural images. In Frontiers in Computational Neuroscience. Conference Abstract: Bernstein Symposium. doi:10.3389/conf.neuro.10.2008.01.116.

[69] Snelbecker, G. E. (1983). Learning theory, instructional theory, and psychoeducational design. New York: McGraw-Hill.

[70] Sweeney, L. (2003). That's AI?: a history and critique of the field. Technical Report, CMU-CS-03-106. Pittsburgh: Carnegie Mellon University, School of Computer Science

[71] Swift, B., Jain, L., White, C., Chandrasekaran, V., Bhandari, A., Hughes, D. A., et al. (2018). Innovation at the Intersection of Clinical Trials and Real-World Data Science to Advance Patient Care. Clin. Transl. Sci. 11, 450–460. doi: 10.1111/cts.12559

[72] SyncedReview (2018) 2018 in Review: 10 AI Failures https://medium.com/syncedreview/2018-in-review-10-ai-failures-c18faadf5983

[73] Walker, S., 2017. Learning theory and behaviour modification. Routledge.

[74] Wang. Y., and Chiew, V. (2010) On the cognitive process of human problem solving, Cognitive Systems Research, Vol.11, Issue 1, Mar, Pages 81-92

[75] Weegar, M.A. and Pacis, D., 2012. A Comparison of two theories of learning-behaviorism and constructivism as applied to face-to-face and online learning. In Proceedings e-leader conference, Manila.

[76] White, R. (2017) Building trust in real-world evidence and comparative effectiveness research: the need for transparency, J. Comp. Eff. Res. 6, 5–7

[77] Wright, L. M., & Leahey, M. (2000). Nurses and families: A guide to family assessment and intervention (3rd ed.. Philadelphia: F. A. Davis

[78] Yampolskiy, R.V. (2018), Artificial Intelligence safety and security, Taylor Francis, London

[79] Yampolskiy, R. V. (2019), "Predicting future AI failures from historic examples", Foresight, Vol. 21 No. 1, pp. 138-152.

[80] Yang, Y., and Bringsnord, S. (forthcoming) Mental metalogic: A new, unifying theory of human and machine reasoning. Erlbaum. Cited in Anderson, J.R., and Lebiere, C. (2003) The Newell Test for a Theory of Cognition, Behavioral and Brain Sciences, Vol.26, pp587-640

[81] Zhao, Y. et al (2017)Exploration of machine learning techniques in predicting multiple sclerosis disease course, PLoS ONE, 12, p. e0174866

[82] Zadeh L.A. (1965), Fuzzy Sets, Information and Control, 8, 338-353, 1965.

An Ontology Driven ESCO LOD Quality Enhancement

Adham Kahlawi
Department of Statistics, Computer Science, Applications
University of Florence, Florence, Italy

***Abstract*—The labor market is a system that is complex and difficult to manage. To overcome this challenge, the European Union has launched the ESCO project which is a language that aims to describe this labor market. In order to support the spread of this project, its dataset was presented as linked open data (LOD). Since LOD is usable and reusable, a set of conditions have to be met. First, LOD must be feasible and high quality. In addition, it must provide the user with the right answers, and it has to be built according to a clear and correct structure. This study investigates the LOD of ESCO, focusing on data quality and data structure. The former is evaluated through applying a set of SPARQL queries. This provides solutions to improve its quality via a set of rules built in first order logic. This process was conducted based on a new proposed ESCO ontology.**

Keywords—ESCO; linked open data; ontology; semantic web; data quality; SPARQL; OWL; metadata

I. Introduction

Labor market governance is one of Europe’s top priorities. Market governance is an important challenge because the job market is a complex network involving many diverse actors. Therefore, the European Commission has proposed European Skills, Competences, Qualifications and Occupations (ESCO)[1] (the multilingual European Skills, Competences, Qualifications, and Occupations classification) as a standard language of work. To enhance its use and reuse, ESCO has published its dataset as Linked Open Data (LOD). Meanwhile, some intelligent services have been provided by the use of LOD like entity search, personalized recommendation and so on [1][2]; Furthermore, the ability to add a language tag to different labels [3], which belongs to one Universal Resource Identifier (URI), enables the use of this system in different countries. For instance, the financial crisis of 2007–2008 increased the rate of unemployment in Europe, especially in Spain where youth unemployment exceeded 50 percent [4]. At the same time in some economic sectors such as engineering and healthcare, companies were not able to find the workforce they need [5]. The EU seeks to reduce this problem by achieving two objectives: 1) helping the jobseekers find a suitable job in another European country, and 2) enabling people to refocus on their careers with a future outlook [6]. Based on this, ESCO was born to help someone who studied in Germany, and lived in Greece to work in Italy by the linked open data that achieve semantic interoperability throughout Europe. Nevertheless, data diffusion is not the only priority to have a good knowledge system on the labor market also data quality has to be assessed. Data quality has always been the focus of researchers' attention for the many challenges it faces [7][8]. Several methodologies have been developed to enhance as well as to assess data quality [9]. For these reasons, any Linked Open Data (LOD) has to consider these aspects before being published.

In order to solve these issues, this study seeks to make the ESCO LOD more structured and more accurate in providing search results.

Section 2 in this article addresses the concept of data quality, data quality dimensions and the related methods of evaluation. Section 3 explains the ESCO structure in details. Section 4 provides a proposal to redesign the structure of ESCO ontology. Section 5 evaluates the new ontology.

II. Linked Open Data and Data Quality

The LOD has been considered as the cornerstone of the semantic web vision and as windows through which data is published in the web. Nowadays there are millions of LOD published in the web [10] at different quality. The data quality is defined as the ability to use and reuse data in a particular application or use case [11]. Data with quality problems might be useful in some cases as long as the quality is within the required range [12]. Nevertheless, it has many challenges. In particular, as explained in [13], the data is published by different providers so that a question of data confidence might be raised. Second, data increases rapidly, making its quality difficult to assess. Third, the level of data quality has been determined from the point of view of the system provider. In fact, when LOD is reused for a different purpose to the initial intention of the provider, certain difficulties are encountered due to the issue of data quality required for the new objective. Data quality has multiple dimensions [14]. In addition, these dimensions range from accessibility to completeness through comprehension. The quality dimensions pose certain challenges [15] such as: a) the issues that the quality of information is dependent solely on the data provider, b) the rapid increase of amount of data makes it more difficult to assess its quality, c) the preparation of the linked open data to be able to reused by third party in a way not expected by the provider, d) the linked open data is a dynamic environment, which requires up-to-date changes to reflect the real world.

Although data quality cannot be assessed with an absolute measurement, LOD can be considered as a useful tool to determine its fitness for reuse.

[1] https://ec.europa.eu/esco/portal

Multiple methodologies have been developed to improve the quality of linked open data; such as: using the statistical distributions to increase the quality of incomplete and noisy Linked Data sets [16]. The authors proposed a method to demonstrate the understandability problems of Resource Description Framework (RDF) data by using the different technologies provided by the semantic web.

The assessment of quality for LOD can be divided into three categories: automated [17], semiautomatic [18], and manual [19]. This article adopts the methodology used in "Test-driven Evaluation of Linked Data Quality" [20] to assess the quality LOD. The method defines some query based text cases implemented with the use of SPARQL (query language for RDF) query templates.

This article focuses on the case of the LOD of the project European Skills, Competences, Qualifications and Occupations (ESCO).

III. ESCO LOD Assessment

ESCO has published its ontology and its LOD. In Fig. 1, the ESCO ontology[2] is depicted while Fig. 2, exhibits the class structure of ESCO LOD that is represented by stardog server[3].

LOD is the new opportunity for sharing and reusing; meanwhile, the ontology forms the main joint of this LOD that weaves the data together [21]. In contrast, comparing these two structures ESCO ontology and ESCO LOD identifies some questions. In order to assess the capability of the current ESCO ontology to being exploited of retrieve valuable information from the related LOD, according to [20] we identified a number of assessment queries.

A. *Resource Description Framework Schema and Web Ontology Language Metadata in ESCO LOD are Missing*

It can be argued that the concepts of class, subclass, data property, object property, and individual lacks a clear definition.

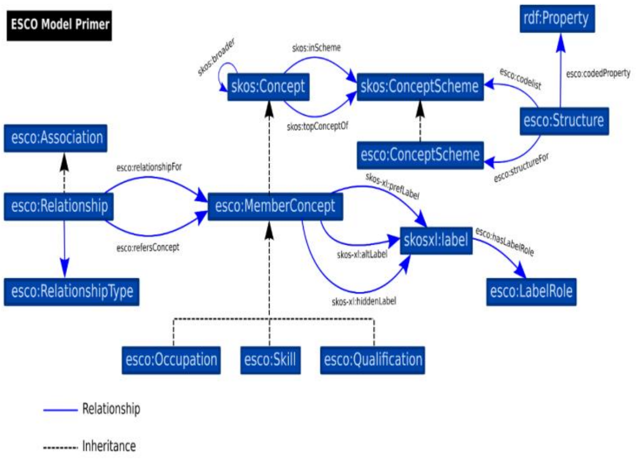
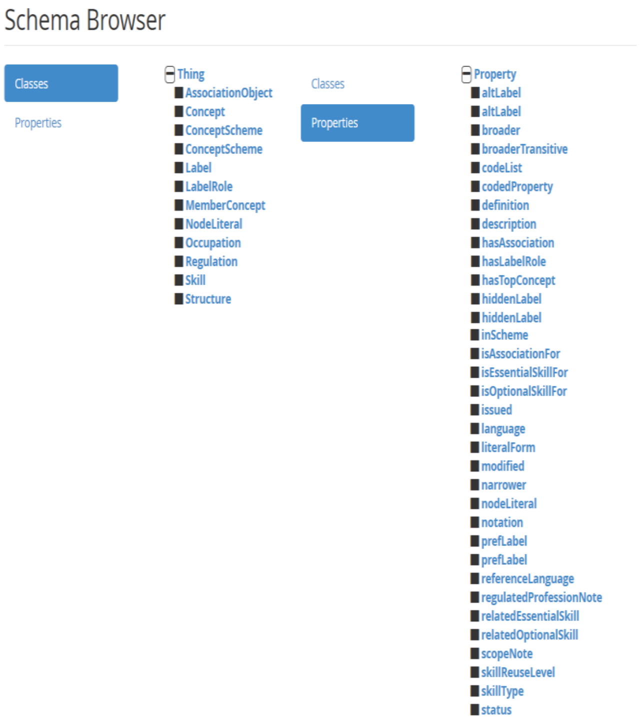


Fig 1. ESCO Ontology.


Schema Browser

Classes
Properties

- Thing
  - AssociationObject
  - Concept
  - ConceptScheme
  - ConceptScheme
  - Label
  - LabelRole
  - MemberConcept
  - NodeLiteral
  - Occupation
  - Regulation
  - Skill
  - Structure

Classes
Properties

- Property
  - altLabel
  - altLabel
  - broader
  - broaderTransitive
  - codeList
  - codedProperty
  - definition
  - description
  - hasAssociation
  - hasLabelRole
  - hasTopConcept
  - hiddenLabel
  - hiddenLabel
  - inScheme
  - isAssociationFor
  - isEssentialSkillFor
  - isOptionalSkillFor
  - issued
  - language
  - literalForm
  - modified
  - narrower
  - nodeLiteral
  - notation
  - prefLabel
  - prefLabel
  - referenceLanguage
  - regulatedProfessionNote
  - relatedEssentialSkill
  - relatedOptionalSkill
  - scopeNote
  - skillReuseLevel
  - skillType
  - status

Fig 2. ESCO LOD Structure.

Binding between two resources to indicate that the first resource is sub concept of the other depends on the two properties of Simple Knowledge Organization System (SKOS) "broader and narrower". Meanwhile, one of these resources or both can be part of a classification. However, this way does not differentiate between a concept that represents a certain level of classification and the individuals contained in this level. To acquire all the skills connected with an occupation is a straightforward task:

SELECT DISTINCT ?skill

where{

?skill rdf:type <http://data.europa.eu/esco/model#Skill>.

?occupation rdf:type

<http://data.europa.eu/esco/model#Occupation>.

?occupation ?property ?skill.

FILTER(?property In

(<http://data.europa.eu/esco/model#relatedEssentialSkill>, <http://data.europa.eu/esco/model#relatedOptionalSkill>))}

In contrast applying a SPARQL query to obtain all the skills which are not connected with an occupation is impossible. In other words, a query as the following one:

SELECT DISTINCT ?skill

WHERE{

?skill rdf:type <http://data.europa.eu/esco/model#Skill>.

?occupation rdf:type <http://data.europa.eu/esco/model#Occupation>.

[2] https://ec.europa.eu/esco/resources/data/static/model/html/model.xhtml
[3] https://www.stardog.com/

```
FILTER NOT EXISTS {?occupation ?property ?skill.}}
```

returns not only skills which are not connected with an occupation, but also all the resources which represent the hierarchical structure of the skill concept.

The benefit of using this metadata is that it facilitates the reuse [22] and supports reasoning in all profiles of Web Ontology Language (OWL). Moreover, since query answering is reduced to OWL-QL query answering, this allows queries to be run over large ontologies [23].

B. Label and Description

The label properties *altLabel*, *hiddenLabel* and *preflabel*, are used to provide a label to a resource. Each property has two namespaces: the first is SKOS, which links the resource to the literal object; the second one is the extension Simple Knowledge Organization (SKOS-XL), which links the resource with one or more resource type *SKOS-XL:Label* which in turn has a "*literalForm*" feature with the same role of SKOS's previous property. However, if the resource contains more than one resource from *SKOS-XL:Label*, each one belongs to label written in a specific language. Additionally, the definition and the description are properties that provide a description to a re-source where the definition property is used only 54 time concurrently with description property. Each resource is collected with one or more resource which in turn has property "*nodeLiteral*" containing a literal object that includes the description with a "*language*" property that indicates the language used to write the description. In case the resource is collected with one resource then the description is written in English. However, if the resource is collected by more than one resource, each one belongs to the description written in a specific language. Consequently, the dataset of ESCO include duplicate information. Therefore, data exploration becomes more difficult and a storage space increases.

C. The Relationship between Skill and Occupation

The relationship between skills and occupation has been built by only two predicates "*relatedEssentialSkill* and *relatedOptionalSkill*". At the same time, the skills in ESCO dataset are divided into two type "skill and knowledge" by a triple that has the skill as subject, skill type as predicate and the type of the skill as an object where each Skill belongs to only one type. The SPARQL query that returns the skills and the knowledge of an occupation, it is very complicated and is written in the following format:

```
SELECT ?essentialskill ?optionalskill ?essentialknowledge
?optionalknowledge

WHERE{

{?essentialskill rdf:type
<http://data.europa.eu/esco/model#Skill>;

<http://data.europa.eu/esco/model#skilltype>
<http://data.europa.eu/esco/skill-type/skill>.

?occupation
<http://data.europa.eu/esco/model#relatedessentialskill>
?essentialskill.}

UNION{?essentialknowledge rdf:type
<http://data.europa.eu/esco/model#Skill>;

<http://data.europa.eu/esco/model#skilltype>
<http://data.europa.eu/esco/skill-type/knowledge>.

 ?occupation
<http://data.europa.eu/esco/model#relatedessentialskill>
?essentialknowledge.}

UNION{?optionalskill rdf:type
<http://data.europa.eu/esco/model#Skill>;

<http://data.europa.eu/esco/model#skilltype>
<http://data.europa.eu/esco/skill-type/skill>.

 ?occupation
<http://data.europa.eu/esco/model#relatedoptionalskill>
?optionalskill.}


UNION{?optionalknowledge rdf:type
<http://data.europa.eu/esco/model#Skill>;

<http://data.europa.eu/esco/model#skilltype>
<http://data.europa.eu/esco/skill-type/knowledge>.

 ?occupation
<http://data.europa.eu/esco/model#relatedoptionalskill>
?optionalknowledge.}

FILTER(?occupation in
(<http://data.europa.eu/esco/occupation/1b4e795d-6e49-4b7b-
bb34-585edfd6eb18>))

}
```

This complexity in query formulation consequent to triples diversity causes slow execution of the SPARQL query [24]. The principal impediment a user faces when trying to apply a query is that he mostly has no information about the LOD underlying structure.

D. Skill and Occupation Structure

The structure of skill and occupation has been discovered within the linked open data of ESCO by applying some query and by using the information represented in class *esco:Structure*.

The occupation structure consists of six levels, the first four levels are based on International Standard Classification of Occupations (ISCO), and the last two levels can be considered as instances of the fourth level. The relation between each level is managed by some predicate like *skos:broader*, *skos:broaderTransitive* and *skos:narrower*. The resources of ESCO classification are generated from type of *skos:Concept*. However, the occupations resources are generated from type *skos:Concept*, *MemberConcept* and *Occupation*.

On the other hand, the skills structure has nothing to do with standard classification and not tied to a consistent classification where the classification branches have different lengths. The first two levels of the classification can be considered as classes and the rest of classification levels can be considered as instances. The relation between one level and another is managed by some predicate like *skos:broader, skos:broaderTransitive* and *skos:narrower.*

All in all, this structure only complicates the data, making it difficult for the user to understand and manipulate.

E. The type of Concept, ConceptScheme and MemberConcept

OWL ontologies and LOD are increasing; thus, the need to give more accurate descriptions of their sources is becoming more necessary [25]. When a general type of class contains sources that only belong to this class or for other classes at the same time cause difficulty to discover their roles and their relations within the linked open data by the user; for example, each resource represents a skill is from *Skill, concept* and *MemberConcept* type; instead, each resource represents a skill reuse level is from Concept type only. SKOS classes can consider them as a representative that establishes an “indirection role’’ between lexical entities and ‘‘real-world’’ but not as a representative of the ‘‘real-world’’ [26].

IV. The Proposed Ontological Model to Reconstruct the ESCO LOD

Nowadays information and systems are growing more rapidly and becoming more complex. As a result, there has to be a method to generate the result of improving the information and the systems with shorter lead-times at less cost [27]. For the semantic data, this method is represented by the rules that define new concepts, relations and metadata which provide a real definition of each resource in the LOD [28] [29][30] All the rules included in appendix “first order logic rules”.

Fig. 3 represents the proposed ontology for ESCO. This model was built by implementing a set of rules written in first order logic. Each set of these rules has a specific task in building the model as follows

A. Classification Building

The model consists of two classifications: one represents the occupation and the other represents the skill. In terms of occupation, the structure is divided into two parts: the first part displays the hierarchical structure represented by rules from 1 to 8, and the second part shows individuals represented by rules from 9 to 16. In terms of skill, the structure is divided into two parts: the first part presents the hierarchical structure represented by rules from 17 to 24, and the second part presents individuals represented by rules from 25 to 44.

B. Give Entities to Different Resources in LOD

The proposed model encompasses classes that did not exist in the ESCO ontology to express the nature and the entity of some the sources that were under general classes. In fact, it can only be identified by relations. The rules between 45 and 54 represent the process of creating new classes and adding individuals to each one.

C. Create the Object Properties of Proposed Ontological Model

The proposed ontological model contains new object properties that represent the relations amongst the new classes. It also contains new relations that describe the relations amongst the existing classes in ESCO ontology in a more accurate manner. Rules 55 to 82 describe the process of establishing these object properties.

D. Stay away from Duplicate Data that Achieve the Same Goal

The article demonstrates that in ESCO LOD has been used the vocabulary of SKOS and the vocabulary of SKOS-XL as noted in the paragraph 3.2. The vocabulary of SKOS-XL is used when is needed to add more information to a label or a description [31]. Nonetheless, the ESCO LOD has not added any other metadata information for this reason, the vocabulary of SKOS-XL has been excluded and only used the vocabulary of SKOS.

V. Evaluation of the Proposed ESCO Ontology

The evaluation of the proposed ontology is based on three criteria:

- The ability to know the contents of the dataset and the mechanism of linking these contents through the ontological schema.

Through the ontological scheme we can understand the following issues: the individuals of class Skill have two different natures; consequently, it can be Skill or knowledge. To be able to perform an occupation, one needs to have some essential skills and knowledge and some optional skills and knowledge. Also to be able to have a skill or a knowledge, one needs to have some essential skills and knowledge and some optional skills and knowledge.

- Preventing information duplication and reducing dataset size.

The ESCO LOD uses two ways to add the labels to a resource as we see before, in spite of the pro-posed ontology use

The direct way to add the labels for a resource accordingly, it prevents the duplicate information and reduce the dataset size by more than three million and half triples.

- Easy retrieval of data through SPARQL queries.

The proposed ontology includes four object properties to connect an occupation or a skill with their essential or optional skills and knowledge. Consequently, it is easy to write a SPARQL query to know which skills or knowledge are essential and which ones are optional to perform an occupation or to obtain a new skill or knowledge.

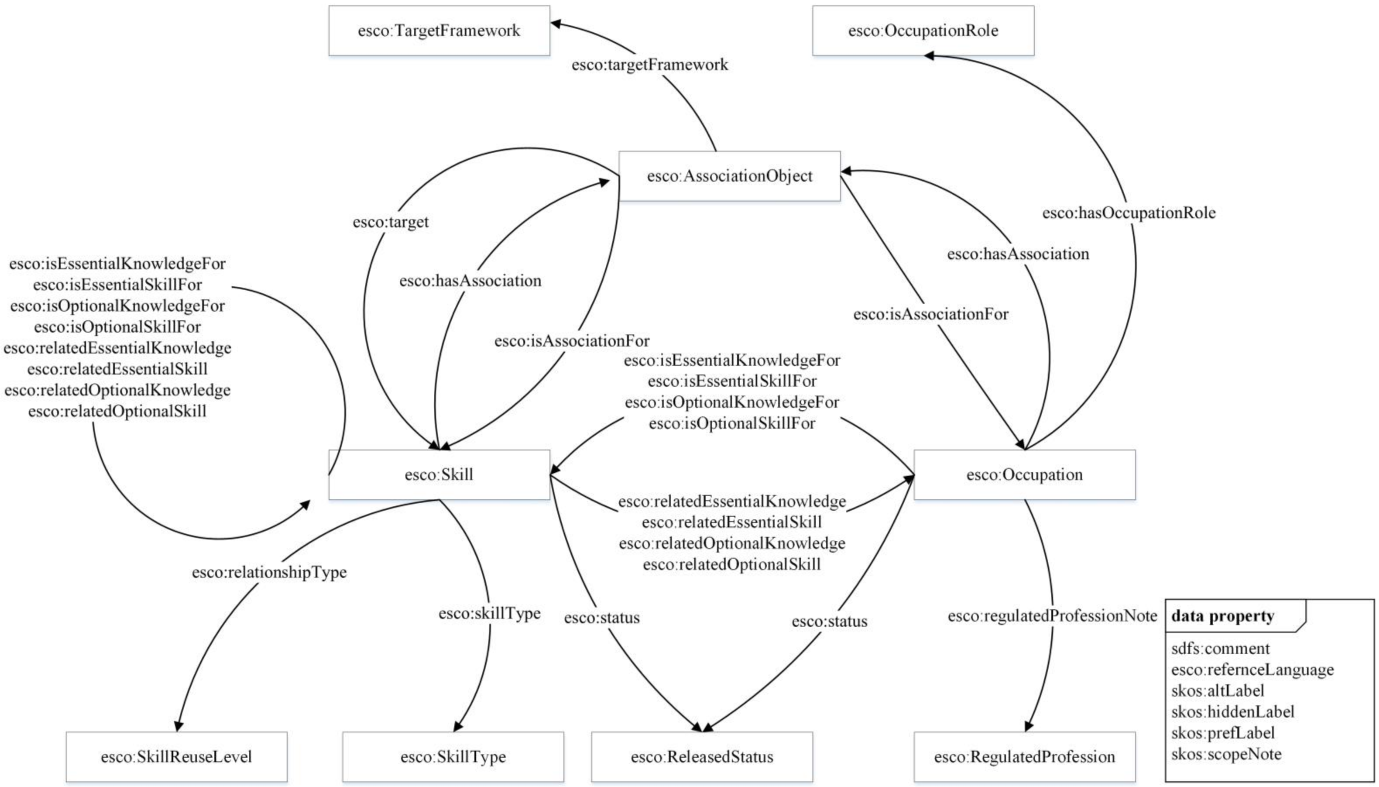


Fig 3. Proposed ESCO Ontology.

A. Getting all Skill which are Not Connected with an Occupation

Applying a SPARQL query to answer this ques-tion depending on a new ESCO ontology is as follows:

```
SELECT DISTINCT ?skill

WHERE{

?skill rdf:type <http://data.europa.eu/esco/model#Skill>, owl:namedindividual.


?occupation rdf:type <http://data.europa.eu/esco/model#Occupation>.


FILTER NOT EXISTS {?occupation ?property ?skill.}}
```

B. Acquiring All Skills and Knowledge of an Occupation

When we add other two properties to represent the relation between skills and occupations in the new ESCO ontology, the query will be more clear and more simple. For instance, get all the skills and knowledge for the occupation "footwear production machine operator" and "footwear designer"

```
SELECT ?essentialskill ?optionalskill ?essentialknowledge ?optionalknowledge

WHERE{

{?occupation <http://data.europa.eu/esco/model#relatedessentialskill> ?essentialskill.}

UNION{?occupation <http://data.europa.eu/esco/model#relatedessentialknowledge> ?essentialknowledge.}

UNION{?occupation <http://data.europa.eu/esco/model#relatedoptionalskill> ?optionalskill.}

UNION{?occupation <http://data.europa.eu/esco/model#relatedoptionalknowledge> ?optionalknowledge.}

FILTER(?occupation in (<http://data.europa.eu/esco/occupation/1b4e795d-6e49-4b7b-bb34-585edfd6eb18>,

<http://data.europa.eu/esco/occupation/06f89f2c-c6e9-40c5-a4a5-0e34d5fbc184>))

}
```

VI. Conclusion

ESCO is one of the most important European projects aimed at modeling the labor market. Its LOD is one of the most qualified LOD for reuse. Thus, it has to be clear and as easy to use as possible. In the proposed ontological model, this study relied on a set of conditions to maintain clarity, such as:

- Non-repetition data
- Using OWL and RDFS to build classifications and to identify each source and whether this source represents a class, individual, object property or data property.

- Determining the dependency of each source for a specific class illustrating the nature of this source.

One of the more significant findings to emerge from this study is that the proposed ontological model could be a pillar of a new version of the ESCO LOD in the coming years since the European Union will adopt this data at the level of all member states. The current study makes several noteworthy contributions to improve the outputs of studies that aim to use ESCO LOD as a tool for search and job matching, career management, and labor market analysis.

The methods used for this study to improve the data quality and data structure of ESCO LOD may be applied to other datasets published as LOD elsewhere in the world.

The following conclusions can be drawn from the present study. The ESCO LOD could be not only one of the most important sources of information for building job applications, but also a basis for a recommendation system for building an effective training system in all member states. This is expected to yield several benefits arising from the advantages of hierarchical structure for classifications of some classes within the data. Another benefit will result from the advantages of horizontal structure arising from relationships between the classes, as well as qualifications issued by private awarding bodies.

References

[1] X. Niu, X. Sun, H. Wang, S. Rong, G. Qi, and Y. Yu, “Zhishi. me-weaving chinese linking open data,” in International Semantic Web Conference, 2011, pp. 205–220.

[2] T. Di Noia, R. Mirizzi, V. C. Ostuni, D. Romito, and M. Zanker, “Linked open data to support content-based recommender systems,” in Proceedings of the 8th international conference on semantic systems, 2012, pp. 1–8.

[3] B. Ell, D. Vrandečić, and E. Simperl, “Labels in the web of data,” in International Semantic Web Conference, 2011, pp. 162–176.

[4] European Commission, Labour Market Developments in Europe. 2012.

[5] European Commission, “European Commission Digital Jobs launches Grand Coalition for,” 2013.

[6] M. Le Vrang, A. Papantoniou, E. Pauwels, P. Fannes, D. Vandensteen, and J. De Smedt, “ESCO: Boosting job matching in Europe with semantic interoperability,” Computer, vol. 47, no. 10, pp. 57–64, 2014.

[7] L. Cai and Y. Zhu, “The challenges of data quality and data quality assessment in the big data era,” Data science journal, vol. 14, 2015.

[8] I. Taleb, H. T. El Kassabi, M. A. Serhani, R. Dssouli, and C. Bouhaddioui, “Big data quality: A quality dimensions evaluation,” in 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), 2016, pp. 759–765.

[9] M. Färber, F. Bartscherer, C. Menne, and A. Rettinger, “Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago,” Semantic Web, vol. 9, no. 1, pp. 77–129, 2018.

[10] K. Jacksi, S. R. Zeebaree, and N. Dimililer, “LOD Explorer: Presenting the Web of Data,” Intl. Journal of Advanced Computer Science and Applications, vol. 9, no. 1, pp. 45–51, 2018.

[11] S. Knight and J. M Burn, “Developing a Framework for Assessing Information Quality on the World Wide Web,” Informing Science Journal, vol. 8, pp. 159–172, 2005.

[12] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, “Quality assessment for linked data: A survey,” Semantic Web, vol. 7, no. 1, pp. 63–93, 2016.

[13] H. H. Ahmed, “Data quality assessment in the integration process of linked open data (LOD),” in 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), 2017, pp. 1–6.

[14] L. L. Pipino, Y. W. Lee, and R. Y. Wang, “Data quality assessment,” Communications of the ACM, vol. 45, no. 4, pp. 211–218, 2002.

[15] C. Batini and M. Scannapieco, Data and Information Quality: Dimensions, Principles and Techniques. Springer International Publishing, 2016.

[16] H. Paulheim and C. Bizer, “Improving the quality of linked data using statistical distributions,” International Journal on Semantic Web and Information Systems (IJSWIS), vol. 10, no. 2, pp. 63–86, 2014.

[17] C. Guéret, P. Groth, C. Stadler, and J. Lehmann, “Assessing linked data mappings using network measures,” in Extended semantic web conference, 2012, pp. 87–102.

[18] A. Flemming, “Quality characteristics of linked data publishing datasources,” Master’s thesis, Humboldt-Universität of Berlin, 2010.

[19] C. Bizer and R. Cyganiak, “Quality-driven information filtering using the WIQA policy framework,” Journal of Web Semantics, vol. 7, no. 1, pp. 1–10, 2009.

[20] D. Kontokostas et al., “Test-driven evaluation of linked data quality,” in Proceedings of the 23rd international conference on World Wide Web, 2014, pp. 747–758.

[21] M. C. Pattuelli, A. Provo, and H. Thorsen, “Ontology building for linked open data: A pragmatic perspective,” Journal of Library Metadata, vol. 15, no. 3–4, pp. 265–294, 2015.

[22] P.-Y. Vandenbussche and B. Vatant, “Metadata recommendations for linked open data vocabularies,” Version, vol. 1, pp. 2011–2012, 2011.

[23] E. Thomas, J. Z. Pan, and Y. Ren, “TrOWL: Tractable OWL 2 reasoning infrastructure,” in Extended Semantic Web Conference, 2010, pp. 431–435.

[24] S. Campinas, T. E. Perry, D. Ceccarelli, R. Delbru, and G. Tummarello, “Introducing RDF graph summary with application to assisted SPARQL formulation,” in 2012 23rd International Workshop on Database and Expert Systems Applications, 2012, pp. 261–266.

[25] A. Kalyanpur, B. Parsia, E. Sirin, and J. Hendler, “Debugging unsatisfiable classes in OWL ontologies,” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 3, no. 4, pp. 268–293, 2005.

[26] T. Baker, S. Bechhofer, A. Isaac, A. Miles, G. Schreiber, and E. Summers, “Key choices in the design of Simple Knowledge Organization System (SKOS),” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 20, pp. 35–49, 2013.

[27] A. Abadi, H. Ben-Azza, and S. Sekkat, “Improving integrated product design using SWRL rules expression and ontology-based reasoning,” Procedia Computer Science, vol. 127, pp. 416–425, 2018.

[28] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz, “Pellet: A practical owl-dl reasoner,” Web Semantics: science, services and agents on the World Wide Web, vol. 5, no. 2, pp. 51–53, 2007.

[29] N. Alaya, M. Lamolle, and S. Ben Yahia, “Multi-Label Based Learning for Better Multi-Criteria Ranking of Ontology Reasoners,” in International Semantic Web Conference, 2017, pp. 3–19.

[30] Z. Quan and V. Haarslev, “A framework for parallelizing OWL classification in description logic reasoners,” arXiv preprint arXiv:1906.07749, 2019.

[31] A. Isaac and E. Summers, “SKOS simple knowledge organization system,” Primer, World Wide Web Consortium (W3C), vol. 44, 2009.

APPENDIX FIRST ORDER LOGIC RULES

A. Classification Building of Occupation

1) Class Hierarchy

Rule (1)

```
∀x,y,z (skos:narrower(x,y) ∧
(relatedEssentialSkill(y,z) ∨
relatedOptionalSkill(y,z)) ∧ ¬
(relatedEssentialSkill(x,z) ∧
relatedOptionalSkill(x,z)) → owl:Class(x)
)
```

Rule (2)

```
∀x,y,z,u (skos:narrower(x,y) ∧
(relatedEssentialSkill(y,z) ∨
relatedOptionalSkill(y,z)) ∧ ¬
(relatedEssentialSkill(x,z) ∧
relatedOptionalSkill(x,z)) ∧
skos:narrower(u,x)   → owl:Class(u) )
```

Rule (3)

```
∀x,y,z,u (skos:narrower(x,y) ∧
(relatedEssentialSkill(y,z) ∨
relatedOptionalSkill(y,z)) ∧ ¬
(relatedEssentialSkill(x,z) ∧
relatedOptionalSkill(x,z)) ∧
skos:narrower(u,x)   →
rdfs:subClassOf(u,x))
```

Rule (4)

```
∀x,y,z,u,s (skos:narrower(x,y) ∧
(relatedEssentialSkill(y,z) ∨
relatedOptionalSkill(y,z)) ∧ ¬
(relatedEssentialSkill(x,z) ∧
relatedOptionalSkill(x,z)) ∧
skos:narrower(u,x) ∧ skos:narrower(s,u)
→ owl:Class(s) )
```

Rule (5)

```
∀x,y,z,u,s (skos:narrower(x,y) ∧
(relatedEssentialSkill(y,z) ∨
relatedOptionalSkill(y,z)) ∧ ¬
(relatedEssentialSkill(x,z) ∧
relatedOptionalSkill(x,z)) ∧
skos:narrower(u,x) ∧ skos:narrower(s,u)
→ rdfs:subClassOf(s,u))
```

Rule (6)

```
∀x,y,z,u,s ,d (skos:narrower(x,y) ∧
(relatedEssentialSkill(y,z) ∨
relatedOptionalSkill(y,z)) ∧ ¬
(relatedEssentialSkill(x,z) ∧
relatedOptionalSkill(x,z)) ∧
skos:narrower(u,x) ∧ skos:narrower(s,u)
∧ skos:narrower(d,s)     →  owl:Class(d))
```

Rule (7)

```
∀x,y,z,u,s ,d (skos:narrower(x,y) ∧
(relatedEssentialSkill(y,z) ∨
relatedOptionalSkill(y,z)) ∧ ¬
(relatedEssentialSkill(x,z) ∧
relatedOptionalSkill(x,z)) ∧
skos:narrower(u,x) ∧ skos:narrower(s,u)
∧ skos:narrower(d,s)     →
rdfs:subClassOf(d,s))
```

Rule (8)

```
∀x,y,z,u,s ,d (owl:class(Occupation) ∧
skos:narrower(x,y) ∧
(relatedEssentialSkill(y,z) ∨
relatedOptionalSkill(y,z)) ∧ ¬
(relatedEssentialSkill(x,z) ∧
relatedOptionalSkill(x,z)) ∧
skos:narrower(u,x) ∧ skos:narrower(s,u)
∧ skos:narrower(d,s)     →
rdfs:subClassOf(Occupation, d))
```

2) adding individuals

Rule (9)

```
∀x,y,z (skos:narrower(x,y) ∧
(relatedEssentialSkill(y,z) ∨
relatedOptionalSkill(y,z)) ∧ ¬
(relatedEssentialSkill(x,z) ∧
relatedOptionalSkill(x,z)) →
owl:NamedIndividual(y) )
```

Rule (10)

```
∀x,y,z (skos:narrower(x,y) ∧
(relatedEssentialSkill(y,z) ∨
relatedOptionalSkill(y,z)) ∧ ¬
(relatedEssentialSkill(x,z) ∧
relatedOptionalSkill(x,z)) →
rdf:type(x,y))
```

Rule (11)

```
∀x,y,z,u (skos:narrower(x,y) ∧
(relatedEssentialSkill(y,z) ∨
relatedOptionalSkill(y,z)) ∧ ¬
(relatedEssentialSkill(x,z) ∧
relatedOptionalSkill(x,z)) ∧
skos:narrower(y,u)  →
owl:NamedIndividual(u) )
```

Rule (12)

```
∀x,y,z,u (skos:narrower(x,y) ∧
(relatedEssentialSkill(y,z) ∨
relatedOptionalSkill(y,z)) ∧ ¬
(relatedEssentialSkill(x,z) ∧
relatedOptionalSkill(x,z)) ∧
skos:narrower(y,u)  → rdf:type(x,u))
```

Rule (13)

```
∀x,y,z,u,d (skos:narrower(x,y) ∧
(relatedEssentialSkill(y,z) ∨
relatedOptionalSkill(y,z)) ∧ ¬
(relatedEssentialSkill(x,z) ∧
relatedOptionalSkill(x,z)) ∧
```

skos:narrower(y,u) ∧ skos:narrower(u,d) → owl:NamedIndividual(d)

Rule (14)

∀x,y,z,u (skos:narrower(x,y) ∧ (relatedEssentialSkill(y,z) ∨ relatedOptionalSkill(y,z)) ∧ ¬ (relatedEssentialSkill(x,z) ∧ relatedOptionalSkill(x,z)) ∧ skos:narrower(y,u) ∧ skos:narrower(u,d) → rdf:type(x,d))

Rule (15)

∀x,y,z,u,d,f (skos:narrower(x,y) ∧ (relatedEssentialSkill(y,z) ∨ relatedOptionalSkill(y,z)) ∧ ¬ (relatedEssentialSkill(x,z) ∧ relatedOptionalSkill(x,z)) ∧ skos:narrower(y,u) ∧ skos:narrower(u,d) ∧ skos:narrower(d,f) → owl:NamedIndividual(f)

Rule (16)

∀x,y,z,u (skos:narrower(x,y) ∧ (relatedEssentialSkill(y,z) ∨ relatedOptionalSkill(y,z)) ∧ ¬ (relatedEssentialSkill(x,z) ∧ relatedOptionalSkill(x,z)) ∧ skos:narrower(y,u) ∧ skos:narrower(u,d) ∧ skos:narrower(d,f) → rdf:type(x,f))

B. Classification Building of Skill

1) Class Hierarchy

Rule (17)

∀x (x = <http://data.europa.eu/esco/skill/8f18f987-33e2-4228-9efb-65de25d03330> → owl:Class(x))

Rule (18)

∀x (x = <http://data.europa.eu/esco/skill/8f18f987-33e2-4228-9efb-65de25d03330> → rdfs:subClassOf(Skill, x))

Rule (19)

∀x,y (x = <http://data.europa.eu/esco/skill/8f18f987-33e2-4228-9efb-65de25d03330> ∧ skos:narrower(x,y) → owl:Class(y))

Rule (20)

∀x,y (x = <http://data.europa.eu/esco/skill/8f18f987-33e2-4228-9efb-65de25d03330> ∧ skos:narrower(x,y) → rdfs:subClassOf(x,y))

Rule (21)

∀x,y,z (x = <http://data.europa.eu/esco/skill/8f18f987-33e2-4228-9efb-65de25d03330> ∧ skos:narrower(x,y) ∧ skos:narrower(y,z) ∧¬Skill(z)→ owl:Class(z))

Rule (22)

∀x,y,z (x = <http://data.europa.eu/esco/skill/8f18f987-33e2-4228-9efb-65de25d03330> ∧ skos:narrower(x,y) ∧¬Skill(z)→ rdfs:subClassOf(y,z))

Rule (23)

∀x,y,z,a (x = <http://data.europa.eu/esco/skill/8f18f987-33e2-4228-9efb-65de25d03330> ∧ skos:narrower(x,y) ∧ skos:narrower(y,z) ∧ skos:narrower(z,a) ∧¬Skill(z) ∧¬Skill(a)→ owl:Class(a))

Rule (24)

∀x,y,z,a (x = <http://data.europa.eu/esco/skill/8f18f987-33e2-4228-9efb-65de25d03330> ∧ skos:narrower(x,y) ∧ skos:narrower(y,z) ∧ skos:narrower(z,a) ∧¬Skill(z) ∧¬Skill(a)→ rdfs:subClassOf(z,a))

2) adding individuals

Rule (25)

∀x,y,z (x = <http://data.europa.eu/esco/skill/8f18f987-33e2-4228-9efb-65de25d03330> ∧ skos:narrower(x,y) ∧ skos:narrower(y,z) ∧ Skill(z)→ owl:NamedIndividual(z))

Rule (26)

∀x,y,z (x = <http://data.europa.eu/esco/skill/8f18f987-33e2-4228-9efb-65de25d03330> ∧ skos:narrower(x,y) ∧ Skill(z)→ rdf:type(y,z))

Rule (27)

∀x,y (skos:narrower(x,y) ∧ Concept(x) ∧ Skill(y) ∧¬Skill(x) → owl:NamedIndividual(y))

Rule (28)

∀x,y (skos:narrower(x,y) ∧ Concept(x) ∧ Skill(y) ∧¬Skill(x) → rdf:type(x,y))

Rule (29)

∀x,y,z (skos:narrower(x,y) ∧ skos:narrower(y,z) ∧ Concept(x) ∧ Skill(y) ∧ Skill(z) ∧¬Skill(x) → owl:NamedIndividual(z))

Rule (30)

```
∀x,y,z (skos:narrower(x,y) ∧
skos:narrower(y,z) ∧ Concept(x) ∧
Skill(y) ∧ Skill(z) ∧¬Skill(x) →
rdf:type(x,z))
```

Rule (31)

```
∀x,y,z,s (skos:narrower(x,y) ∧
skos:narrower(y,z) ∧ skos:narrower(z,s) ∧
Concept(x) ∧ Skill(y) ∧ Skill(z) ∧
Skill(s) ∧¬Skill(x) →
owl:NamedIndividual(s))
```

Rule (32)

```
∀x,y,z,s (skos:narrower(x,y) ∧
skos:narrower(y,z) ∧ skos:narrower(z,s) ∧
Concept(x) ∧ Skill(y) ∧ Skill(z) ∧
Skill(s) ∧¬Skill(x) → rdf:type(x,s))
```

Rule (33)

```
∀x,y,z,s,a (skos:narrower(x,y) ∧
skos:narrower(y,z) ∧ skos:narrower(z,s) ∧
skos:narrower(s,a) ∧ Concept(x) ∧
Skill(y) ∧ Skill(z) ∧ Skill(s) ∧ Skill(a)
∧¬Skill(x) → owl:NamedIndividual(a))
```

Rule (34)

```
∀x,y,z,s,a (skos:narrower(x,y) ∧
skos:narrower(y,z) ∧ skos:narrower(z,s) ∧
skos:narrower(s,a) ∧ Concept(x) ∧
Skill(y) ∧ Skill(z) ∧ Skill(s) ∧ Skill(a)
∧¬Skill(x) → rdf:type(x,a))
```

Rule (35)

```
∀x,y,z,s,a,b (skos:narrower(x,y) ∧
skos:narrower(y,z) ∧ skos:narrower(z,s) ∧
skos:narrower(s,a) ∧ skos:narrower(a,b) ∧
Concept(x) ∧ Skill(y) ∧ Skill(z) ∧
Skill(s) ∧ Skill(a) ∧ Skill(b) ∧¬Skill(x)
→ owl:NamedIndividual(b))
```

Rule (36)

```
∀x,y,z,s,a,b (skos:narrower(x,y) ∧
skos:narrower(y,z) ∧ skos:narrower(z,s) ∧
skos:narrower(s,a) ∧ skos:narrower(a,b) ∧
Concept(x) ∧ Skill(y) ∧ Skill(z) ∧
Skill(s) ∧ Skill(a) ∧ Skill(b) ∧¬Skill(x)
→ rdf:type(x,b))
```

Rule (37)

```
∀x,y,z,s,a,b,c (skos:narrower(x,y) ∧
skos:narrower(y,z) ∧ skos:narrower(z,s) ∧
skos:narrower(s,a) ∧ skos:narrower(a,b) ∧
skos:narrower(b,c)  ∧ Concept(x) ∧
Skill(y) ∧ Skill(z) ∧ Skill(s) ∧ Skill(a)
∧ Skill(b) ∧ Skill(c) ∧¬Skill(x) →
owl:NamedIndividual(c))
```

Rule (38)

```
∀x,y,z,s,a,b,c (skos:narrower(x,y) ∧
skos:narrower(y,z) ∧ skos:narrower(z,s) ∧
skos:narrower(s,a) ∧ skos:narrower(a,b) ∧
skos:narrower(b,c)  ∧ Concept(x) ∧
Skill(y) ∧ Skill(z) ∧ Skill(s) ∧ Skill(a)
∧ Skill(b) ∧ Skill(c) ∧¬Skill(x) →
rdf:type(x,c))
```

Rule (39)

```
∀x,y,z,s,a,b,c,d (skos:narrower(x,y) ∧
skos:narrower(y,z) ∧ skos:narrower(z,s) ∧
skos:narrower(s,a) ∧ skos:narrower(a,b) ∧
skos:narrower(b,c) ∧ skos:narrower(c, d)
∧ Concept(x) ∧ Skill(y) ∧ Skill(z) ∧
Skill(s) ∧ Skill(a) ∧ Skill(b) ∧ Skill(c)
∧ Skill(d) ∧¬Skill(x) →
owl:NamedIndividual(d))
```

Rule (40)

```
∀x,y,z,s,a,b,c,d (skos:narrower(x,y) ∧
skos:narrower(y,z) ∧ skos:narrower(z,s) ∧
skos:narrower(s,a) ∧ skos:narrower(a,b) ∧
skos:narrower(b,c) ∧ skos:narrower(c, d)
∧ Concept(x) ∧ Skill(y) ∧ Skill(z) ∧
Skill(s) ∧ Skill(a) ∧ Skill(b) ∧ Skill(c)
∧ Skill(d) ∧¬Skill(x) → rdf:type(x,d))
```

3) Individuals unclassified

Rule (41)

```
∀x,y,z ( Concept(x) ∧ Skill(y) ∧ Skill(z)
∧¬Skill(x) ∧ ¬ skos:broader(y,x) ∧ ¬
skos:narrower(y,z) →
owl:NamedIndividual(y))
```

Rule (42)

```
∀x,y,z ( Concept(x) ∧ Skill(y) ∧ Skill(z)
∧¬Skill(x) ∧ ¬ skos:broader(y,x) ∧ ¬
skos:narrower(y,z) →
rdf:type(unclassified,y))
```

Rule (43)

```
∀x,y,z,s ( Concept(x) ∧ Skill(y) ∧
Skill(z) ∧¬Skill(x) ∧ ¬ skos:broader(y,x)
∧ skos:narrower(y,z) ∧ skos:broader(z,y)
∧ ¬ skos:broader(z,s)  →
owl:NamedIndividual(y))
```

Rule (44)

```
∀x,y,z,s ( Concept(x) ∧ Skill(y) ∧
Skill(z) ∧¬Skill(x) ∧ ¬ skos:broader(y,x)
∧ skos:narrower(y,z) ∧ skos:broader(z,y)
∧ ¬ skos:broader(z,s)  →
rdf:type(unclassified,z))
```

C. Class SkillReuseLevel

Rule (45)

```
∀x,y (Skill(x) ∧ Concept(y) ∧
skillReuseLevel(x,y) →
owl:NamedIndividual(y))
```

Rule (46)

```
∀x,y (Skill(x) ∧ Concept(y) ∧
skillReuseLevel(x,y) →
rdf:type(SkillReuseLevel ,y))
```

D. *Class SkillType*

Rule (47)

```
∀x,y (Skill(x) ∧ Concept(y) ∧
skillType(x,y) → owl:NamedIndividual(y))
```

Rule (46)

```
∀x,y (Skill(x) ∧ Concept(y) ∧
skillType(x,y) → rdf:type(SkillType ,y))
```

E. *Class ReleasedStatus*

Rule (47)

```
∀x,y ((Skill(x) ∨ Occupation(x)) ∧
purl:status(x,y) →
owl:NamedIndividual(y))
```

Rule (48)

```
∀x,y ((Skill(x) ∨ Occupation(x)) ∧
purl:status(x,y) →
rdf:type(ReleasedStatus ,y))
```

F. *Class RegulatedProfession*

Rule (49)

```
∀x,y ((Concept(y) ∧ Regulation(y)) ∧
Occupation(x) ∧
regulatedProfessionNote(x,y) →
owl:NamedIndividual(y))
```

Rule (50)

```
∀x,y ((Concept(y) ∧ Regulation(y)) ∧
Occupation(x) ∧
regulatedProfessionNote(x,y) →
rdf:type(RegulatedProfession ,y))
```

G. *Class OccupationRole*

Rule (51)

```
∀x,y,z (Occupation(x) ∧ LabelRole(y) ∧
Label(z) ∧ altLabel(x,z) ∧
hasLabelRole(z,y) →
owl:NamedIndividual(y))
```

Rule (52)

```
∀x,y,z (Occupation(x) ∧ LabelRole(y) ∧
Label(z) ∧ altLabel(x,z) ∧
hasLabelRole(z,y) →
rdf:type(OccupationRole ,y))
```

H. *Class AssociationObject*

Rule (53)

```
∀x (AssociationObject (x) →
owl:NamedIndividual(x))
```

Rule (52)

```
∀x (AssociationObject (x)
→rdf:type(AssociationObject ,x))
```

I. *Class TargetFramework*

Rule (53)

```
∀x,y (AssociationObject(x) ∧
ConceptScheme(y) ∧ targetFramework(x,y) →
owl:NamedIndividual(y))
```

Rule (54)

```
∀x,y (AssociationObject(x) ∧
ConceptScheme(y) ∧ targetFramework(x,y) →
rdf:type(TargetFramework ,y))
```

J. *Object property*

1) *Relatedessentialskill and isEssentialSkillFor*

Rule (55)

```
∀x,y (Skill(x) ∧ Occupation(y) ∧
relatedEssentialSkill(y,x) ∧
skillType(x,skill) →
esco:relatedEssentialSkill(y,x))
```

Rule (56)

```
∀x,y (Skill(x) ∧ Occupation(y) ∧
relatedEssentialSkill(y,x) ∧
skillType(x,skill) →
esco:isEssentialSkillFor(x,y))
```

Rule (57)

```
∀x,y (Skill(x) ∧ Skill(y) ∧
relatedEssentialSkill(y,x) ∧
skillType(x,skill) →
esco:relatedEssentialSkill(y,x))
```

Rule (58)

```
∀x,y (Skill(x) ∧ Skill(y) ∧
relatedEssentialSkill(y,x) ∧
skillType(x,skill) →
esco:isEssentialSkillFor(x,y))
```

2) *Related essential knowledge and Isessentialknowledgefor*

Rule (59)

```
∀x,y (Skill(x) ∧ Occupation(y) ∧
relatedEssentialSkill(y,x) ∧
skillType(x,knowledge) →
esco:relatedEssentialKnowledge (y,x))
```

Roule (60)

```
∀x,y (Skill(x) ∧ Occupation(y) ∧
relatedEssentialSkill(y,x) ∧
skillType(x,knowledge) →
esco:isEssentialKnowledgeFor(x,y))
```

Rule (61)

```
∀x,y (Skill(x) ∧ Skill(y) ∧
relatedEssentialSkill(y,x) ∧
skillType(x,knowledge) →
esco:relatedEssentialKnowledge (y,x))
```

Roule (62)

```
∀x,y (Skill(x) ∧ Skill(y) ∧
relatedEssentialSkill(y,x) ∧
```

skillType(x,knowledge) →
esco:isEssentialKnowledgeFor(x,y))

3) relatedOptionalSkill and isOptionalSkillFor

Rule (63)

∀x,y (Skill(x) ∧ Occupation(y) ∧
relatedOptionalSkill(y,x) ∧
skillType(x,skill) →
esco:relatedOptionalSkill(y,x))

Roule (64)

∀x,y (Skill(x) ∧ Occupation(y) ∧
relatedOptionalSkill(y,x) ∧
skillType(x,skill) →
esco:isOptionalSkillFor(x,y))

Rule (65)

∀x,y (Skill(x) ∧ Skill(y) ∧
relatedOptionalSkill(y,x) ∧
skillType(x,skill) →
esco:relatedOptionalSkill(y,x))

Rule (66)

∀x,y (Skill(x) ∧ Skill(y) ∧
relatedOptionalSkill(y,x) ∧
skillType(x,skill) →
esco:isOptionalSkillFor(x,y))

4) relatedOptionalKnowledge and isOptionalKnowledgeFor

Rule (67)

∀x,y (Skill(x) ∧ Occupation(y) ∧
relatedOptionalSkill(y,x) ∧
skillType(x,knowledge) →
esco:relatedOptionalKnowledge (y,x))

Rule (68)

∀x,y (Skill(x) ∧ Occupation(y) ∧
relatedOptionalSkill(y,x) ∧
skillType(x,knowledge) →
esco:isOptionalKnowledgeFor(x,y))

Rule (69)

∀x,y (Skill(x) ∧ Skill(y) ∧
relatedOptionalSkill(y,x) ∧
skillType(x,knowledge) →
esco:relatedOptionalKnowledge (y,x))

Rule (70)

∀x,y (Skill(x) ∧ Skill(y) ∧
relatedOptionalSkill(y,x) ∧
skillType(x,knowledge) →
esco:isOptionalKnowledgeFor(x,y))

5) status

Rule (71)

∀x,y (Skill(x) ∧ ReleasedStatus(y) ∧
purl:status(x, released) →
esco:status(x,y))

Rule (72)

∀x,y (Occupation(x) ∧ ReleasedStatus(y) ∧
purl:status(x, released) →
esco:status(x,y))

6) regulatedProfessionNote

Rule (73)

∀x,y ((Concept(y) ∧ Regulation(y)) ∧
Occupation(x) ∧
regulatedProfessionNote(x,y) →
esco:regulatedProfessionNote(x,y)

7) skillType

Rule (74)

∀x,y (Skill(x) ∧ Concept(y) ∧
skillType(x,y) → esco:skillType(x,y))

8) relationshipType

Rule (75)

∀x,y (Skill(x) ∧ Concept(y) ∧
skillReuseLevel(x,y) →
esco:relationshipType(x,y))

9) hasOccupationRole

Rule (76)

∀x,y,z(Occupation(x) ∧ LabelRole(y) ∧
Label(z) ∧ altLabel(x,z) ∧
hasLabelRole(z,y) →
esco:hasOccupationRole(x ,y))

10) targetFramework

Rule (77)

∀x,y (AssociationObject(x) ∧
ConceptScheme(y) ∧ targetFramework(x,y) →
esco:targetFramework(x ,y))

11) hasAssociation and isAssociationFor

Rule (78)

∀x,y (Occupation(x) ∧ AssociationObject
(y) ∧ hasAssociation (x,y) →
esco:hasAssociation(x ,y))

Rule (79)

∀x,y (Occupation(x) ∧ AssociationObject
(y) ∧ has Association (x,y) →
esco:isAssociationFor(y,x))

Rule (80)

∀x,y (Skill(x) ∧ AssociationObject (y) ∧
hasAssociation (x,y) →
esco:hasAssociation(x ,y))

Rule (81)

∀x,y (skill(x) ∧ AssociationObject (y) ∧
has Association (x,y) →
esco:isAssociationFor(y,x))

12) Target

Rule (82)

∀x,y (skill(x) ∧ AssociationObject (y) ∧
target(y,x) → esco:target(y,x))

Implementation of a Proof of Concept for a Blockchain-based Smart Contract for the Automotive Industry in Mauritius

Keshav Luchoomun[1], Sameerchamd Pudaruth[2], Somveer Kishnah[3]
Faculty of Information
Communication and Digital Technologies
University of Mauritius

***Abstract*—In recent years, there has been a growth of interest in the blockchain technology across a wide range of industries. Blockchain technology has the potential to transform the way businesses operate especially in the automotive industry. The distributed infrastructure and the secure nature of the blockchain technology encourages trust among businesses and consumers. In Mauritius, the automotive industry is facing challenges such as tampering of vehicle information, falsification of mileage and poor traceability which leads to a lack of trust from customers. In this work, an implementation of a proof of concept (POC) for a blockchain-based smart contract application has been proposed and implemented to mitigate these challenges. The automotives use cases: (a) vehicle importation; and (b) vehicle sale and registration have been implemented in the IBM blockchain platform which provides a secure and transparent way to invoke transactions. Finally, the performance and benefits of the Hyperledger Fabric vehicle application have been assessed based on transparency, security, traceability and efficiency.**

Keywords—Blockchain; smart contract; hyperledger fabric; vehicles; Mauritius

I. Introduction

Blockchain and smart contracts are terms that have constantly been the subject of heated debates across a wide range of industries over the last few years. Blockchain has moved into the mainstream development tools and technologies, with a wide variety of organisations that are keen to explore how it can be implemented for their businesses [1][2]. Smart contract is another term which has received much industry attention. The concept of smart contract was first coined by an American computer scientist, Nick Szabo, who proposed a smart contract to embed contractual clauses based on events such as time or transactions [3]. A smart contract is similar to physical legal contracts in that it is an agreement made between two different parties to finalise the terms of one or more transactions [4].

For decades, both private individuals and businesses have been facing the same dilemma when purchasing a used vehicle. How can the buyer ensure that a vehicle is really in the exact conditions as described by the seller? How can the buyer have trust when buying a pre-owned vehicle without having to worry about its history, mileage, service history, vehicle inspection history and plenty of other factors critical when assessing a vehicle's condition?

Mileage is one of the most significant parameters when assessing the condition of a pre-owned vehicle. It is probably the most widely acknowledged indicator of how 'used' the vehicle is, and therefore has the most significant effect on the vehicle's valuation and the price the buyer pays. Low mileage equals to higher prices and high mileage equals to lower prices. When mileage is such as important indicator of how used the vehicle is, it is critical that the odometer reading is correct. Unfortunately, the vehicle mileage is sometimes tampered with in some manner to conceal this high mileage. This uncertainty has always plagued the used vehicle market. The consequences of tampered vehicles in the same category as 'straight' vehicles have led to unfair and misleading prices. According to an article published by the Défi Media Group, the car industry was shaken by scandals in which odometers were tampered with. The authorised dealer of this car model in Mauritius announced several more cases in which odometer readings were different from those received from the headquarter [5]. Moreover, the process of purchasing vehicle has always been cumbersome and time consuming where multiple parties/authorities are involved and also poses a risk of information manipulation, data duplication and additional transactional costs. One of the major problems plaguing the automotive industry is the sale of counterfeit parts.

The blockchain technology can change the way the automotive industry functions. Blockchain's application will become a part of everyday life across many industries. With the automotive landscape evolving at an exponential rate, it is only fitting that blockchain will find its way into the industry in the early stages of its development. The broad purposes of this paper are: (a) achieve transparency – transaction histories of the vehicle are becoming more transparent with blockchain technology where a distributed ledger is shared among all network participants, (b) enhanced security – blockchain provides an encryption mechanism for approving transactions, (c) improved traceability – in the automotive industry, the purchase of a vehicle needs to undergo a chain of processes and it can be very expensive and cumbersome to trace an item from its origin, (d) increased efficiency and speed – the current system involves lots of paper processes and therefore it is time-consuming, prone to human mistakes and may require mediation from third-parties, (e) reducing costs – for the majority of businesses, cost reductions are a priority. With the implementation of the blockchain technology in the automotive industry, third parties or middleman will no longer be required. The trust is gained from the data on the blockchain, knowing that the ledger is secure, error-free, accurate and immutable.

For this paper, two use cases have been considered: (a) importation of a vehicle from the car manufacturer to the showroom (b) purchase of a vehicle. The application will be a proof of concept to allow the participants in the network to perform transactions in the blockchain platform. Only the digital asset of the blockchain technology has been considered.

The remainder of this paper is organised as follows. Section 2 introduces to literature review on blockchain technologies. The methodology adopted for this study is given in Section 3. Evaluation of the existing system is presented in Section 4. Section 5 shows the proposed system of Hyperledger Fabric vehicle. Section 6 presents the evaluation performance of the proposed system. The conclusion is given in Section 7.

II. Literature Review

The automotive industry of the future will be disruptive from that of today. Blockchain has the potential to play a major role in underpinning the industry transformation for the future. In this section, a systematic study has been carried out to help shape our understanding on the application of blockchain and smart contract to fit business particular use cases.

The first business endeavor that used a blockchain-based ledger was presented in the white-paper published in 2008 by Satoshi Nakamoto describing a system for electronic cash [6] where the Bitcoin technology started. Blockchain is divided as public and private ledgers. Public ledger is accessible to the public domain and everybody can add blocks to them. On the other hand, private ledger only includes a selected group of people [7]. To make a blockchain viable, a single chain is needed, and this is achieved by an agreement of the majority of participants in the network which is known as consensus.

For the automotive industry to implement blockchain technology, it must meet the following requirements [8]:

- Is a business network involved?
- Is consensus used to validate transactions?
- Is an audit trail, or provenance required?
- Must the record of transactions be immutable or tamper proof?

If the automotive business use case answered the first and at least another criterion positively, then it would benefit from the blockchain technology. The network can be designed across organisations or within an organisation.

Smart contract represents the second generation in blockchain technology from a financial transaction protocol to an all-purpose utility. A smart contract is a piece of code that can be executed on a blockchain. The purpose of a smart contract is to enforce the terms and conditions in the agreement. The idea of smart contracts is not new. It was first proposed by Nick Szabo in 1994 [9]. A smart contract can also be considered as an asset manager which controls the allocation of digital assets to the different parties and/or participants in a blockchain network [10]. A smart contract which is self-contained and does not require any external information to be executed is called a deterministic smart contract while one which require inputs from external sources is called a non-deterministic smart contract [11]. Writing smart contracts require both business and technical skills as any bugs can lead to highly unexpected outputs. This can lead to huge business losses or loss in business reputation, which can have negative long-term consequences [12]. There are several payment platforms which are currently in use and which use smart contract. Bitcoin and Ethereum are too such cryptocurrencies. Smart contracts for bitcoin are not Turing complete and supports only monetary transactions. On the other hand, Ethereum is build using Solidity, which is a Turing complete machine. Solidity is a new language designed especially for Ethereum. It has support for loading code from another address and this enables the creation of libraries (pieces of reusable software code) [13].

Hyperledger Fabric is the first blockchain-based system that supports the use of conventional programming languages for building smart contracts and enable complex data queries [14]. This mechanism gives an edge to Hyperledger Fabric and provides possibilities of using smart contracts in enterprise systems [15][16]. Blockchain developer uses Fabric SDK to code the application and smart contract. A registered user interacts with the application by sending an INVOKE order or a QUERY for requesting information. All participants must be registered in the system to have access to membership services. The content of each transaction is encrypted to ensure that only the intended participants can see the content. All transactions are secured, private and confidential. Fabric can only be updated by consensus of the peers. The events are structured as transactions and shared among the different participants. Fabric provides three distinct roles: (a) a committer peer who is responsible for committing transactions, maintaining the ledger and its state, (b) an endorsing peer who receives a transaction proposal for endorsement and it responds by either giving or rejecting endorsement, (c) an ordering peer who approves the new transactions and add them to the ledger.

Many blockchain initiatives are being adopted by car manufacturers around the world. For example, Ford has implemented a blockchain-based system to control the supply chain of cobalt and to ensure that children are not working as labourers anywhere along this supply chain [17]. Volkswagen has built a tracking system to ensure that odometers are not being manipulated to produce deceptive mileage values [18]. Hyundai is working with IBM and its cloud-based AI to improve its supply chain and payment systems [19][20]. In this work, we intend to leverage on the blockchain technology to implement a new system for the automotive ecosystem in Mauritius which will facilitate all processes, reduce cost for all stakeholders and improve customer experience.

III. Methodology

The research in this paper has been conducted as a case study and data has been collected by focusing on business operations and relevant stakeholders. The purpose of this study is to analyse the performance, efficiency and identify flaws when purchasing a vehicle. Subsequently, interviews will be conducted where customised questions will be set to different stakeholders of the automotive business such as buyers and sellers, senior officers from vehicle showroom, marketing, sales and finance, supply chain department of the automotive industry. Information will also be collected through questionnaire-based survey. The research questions identified for this paper are:

Research Question 1: To what extent the buyers/consumers and car dealers/sellers are aware of the benefits of blockchain and smart contract?

This research question is based on case studies where the employees, especially in sales department from the car showroom Ginza Motors has been interviewed to analyse the actual situation regarding the use and implementation of blockchain smart contracts in the automotive industry. The interviews will help to gain knowledge and identifies the flaws in the vehicle purchase process. Moreover, a survey has been conducted to target the buyers and this has showed that a great challenge in the implementation of the proposed framework is to gain the consumer trust and ease in adopting the proposed system. Hence, this question can be answered by analysing the data that have been collected through interviews, on-site observations and questionnaires.

Research Question 2: How can blockchain improve the vehicle purchase process to be more secure, reliable and trustworthy? Are buyers willing to purchase a vehicle that has its information stored in Blockchain with the ease of mind that the purchase they are making is accurate and error-free. How the proposed framework will improve quality (customer's trust, accuracy of vehicle's condition, etc.) in the automotive industry in Mauritius?

This research question can be answered after the proposed framework has been implemented. The collected data has been analysed and interpreted to measure the business success in adopting blockchain and a smart contract solution.

Research Question 3: Has the proposed framework mitigated the barriers or challenges that the automotive industry is facing?

This research question will help to understand the barriers and factors in adopting the blockchain framework in the real world. An in-depth study has been conducted to identify the weaknesses of the automotive industry and based on the data collected through interviews, survey and questionnaires, a framework has been devised to mitigate these weaknesses.

The results from 62 respondents for an online questionnaire survey on Buyers' Experience in purchasing vehicle are discussed as follows.

Q1: To what extent are you aware of the procedures regarding the purchase of vehicle?

We observed that 69.3% of the respondents are not knowledgeable about the procedures involved in vehicle registration and purchase. The main reasons are (a) little information are available online and further assistance/personnel is required to gain more understanding on the process, (b) various authorities are involved in the vehicle registration and purchase and information needs to be collected from different sources. Sometimes, a fraction of the information collected may not be accurate and thus the buyers take more time to verify and validate those information where most of the time the buyers remains doubtful about the process and hesitate to proceed with the purchase. Thanks to the introduction of blockchain technology, this problem can be mitigated, and a smooth and seamless way of processing vehicle registration and purchase is offered to the consumers.

Q2: Do you verify the registration book (horsepower) or other documents before purchasing a vehicle?

We observed that 96.8% of the respondents answered 'Yes' and we can easily state that most of the buyers are meticulous about the vehicle's conditions (mileage, horsepower and other documents) that they are purchasing. These factors determine the cost of the vehicle, durability and maintenance cost. Unfortunately, Mauritius has witnessed many scandals over the years in terms of falsification of horsepower as related in the Défi Media Group on 'Car Traffic. Hence, the adoption of blockchain technology will mitigate this issue as the information of the vehicle will be immutable and each authority in the network will have a clone of the information across the blockchain ecosystem.

Q3: To what extent, do you have trust on the conditions (mileage, authenticity of parts) of the vehicle as described by the sale person?

The most challenging part in purchasing a vehicle is to verify that the vehicle's conditions as described by the salesperson are accurate. A relationship of trust between the buyers and sellers is created. We observed that none of the respondents have chosen options 'Confident' or 'Very Confident'. We can easily state that 100% of the respondents have had an unpleasant experience or remain doubtful when purchasing a vehicle. Thanks to the introduction of the blockchain technology, the vehicle's conditions cannot be tampered, and the authenticity of the parts will remain accurate. Hence, buyers will be able to view a history of the vehicle and parts. As a result, the blockchain technology will create a strong bonding of trust between the buyers and the car dealers.

Q4: How long have you been through a vehicle registration process?

The vehicle registration process is very lengthy. 51.6% of the respondents have waited above 4 weeks to have vehicle ownership and 35.5% of the respondents have waited between 2-4 weeks. The main reason for this lengthy process is because the process involves a lot of paperwork. It is therefore time-consuming and error prone. These transactions can potentially be completed faster and in less cost via the blockchain technology.

Q5: In the vehicle registration process, which step(s) do you find it difficult or time consuming?

We observed that a high percentage of the respondents find it challenging to verify the authenticity of a vehicle's documents, vehicle lease process and apply for insurance. The reason is because in each step different authorities are involved and thus the buyers need to be involved with each authority for documents verification, validation and submission. Moreover, the buyers constantly need to be in communication with each authority to determine the state of the process. As a matter of fact, the steps mentioned above are processed as a chain of responsibility, in other words, verification of authenticity of documents needs to be completed prior to application of vehicle loan. With the introduction of blockchain technology, the authorities/nodes communicate with each other to inform the network that Job A (verification of vehicle's documents) from Node A has been completed and Node B can start Job B (application of car loan) immediately.

Q6: To what extent, are you aware of the benefits of the blockchain technology?

Most of the respondents are not fully versed with the blockchain technology. Hence, workshops, training and campaigns need to be organised to create awareness to the car dealers and public on the blockchain technology. Another way is to design the website or application adhering to the following principles: user-friendly, effective, aesthetic and informative, ease of navigation. In this way, the users/buyers will gain knowledge about the blockchain technology in less time.

Q7: To what extent, are you willing to trust and purchase your vehicle through the blockchain technology?

The majority of the respondents are willing to move away from the traditional system and adopt blockchain technology to purchase their vehicle as the blockchain technology can solve the problem in the existing traditional system by (a) achieving transparency where the vehicle data is more accurate, consistent and no fraud or falsification of horsepower or mileage. Moreover, the authenticity of vehicle's parts can be traced from its origin; (b) increasing efficiency and speed of vehicle registration process to be faster, efficient and eliminate the heavy paper works; (c) reducing cost where buyers will not require paying additional costs.

IV. Evaluation of the Existing System

Registration of vehicles in Mauritius is a very cumbersome process. This is because multiple parties are involved. This creates the risk of data duplication, tampering, manipulation and mismanagement. The process for importation of vehicles in the traditional system is as follows:

1) Car dealer (showroom) needs to have the following prior to importation of vehicles.

a) Clearance from the Ministry of Commerce and Consumer Protection.

b) Importation Permit

c) Bill of Landing

2) Car dealer (showroom) orders vehicles form Manufacturer. Car Dealer can access the vehicle information such as make, model, variant, engine number, chassis number, etc.

3) Manufacturer displays the vehicles in an Auction where there are other suppliers.

4) Car dealer chooses vehicles and makes payments.

5) Vehicles are shipped to Mauritius and remain under MRA (Bond) Custody.

6) Car Dealer needs to complete the following:

a) Register vehicle chassis number (VCN)

b) Duty payment

c) Insurance payment

The following is the process of the traditional system when a customer purchases a vehicle from a car dealer.

1) Customer visits the showroom, chooses vehicle model and request for a test drive.

2) Customer requests for a quotation with vehicle specifications and informs whether purchase will be duty paid or duty free and whether by cash or through leasing.

3) Duty free confirmation letter required from Registrar general for government official or Mauritius Revenue Authority (MRA) for returning resident.

4) The quotation must consist of the showroom price as per duty payable, registration fees, road tax and horsepower fees.

5) Customer confirms purchase, effects a deposit of 15% or more and provides the following documents: photocopies of NIC recto/verso, proof of address (utility bill Central Water Authority, Central Electricity Board, Mauritius telecom), driving license, Payments above ₹ 500 000 is done through office cheque or bank transfer.

6) If purchase is through leasing, the customer has to provide a pay slip, a bank statement for the last 3 months and then waits for approval and the lease documents.

7) Vehicle is sent to workshop for verification, servicing and valeting.

8) Showroom issues sales invoice and customer effects remaining payment.

9) Sales deed document with vehicle specifications is signed by customer and authorised showroom representative.

*10)*Customer and showroom agree on delivery date.

*11)*Sales document is sent for registration with Registrar department.

*12)*Vehicle is insured.

*13)*Vehicle is sent to the National Transport Authority (NTA) for verification, registration and horsepower is issued.

*14)*If vehicle purchase is through leasing a lien is inscribed on the vehicle and horsepower updated accordingly at NTA.

*15)*Road license and insurance vignette is affixed on windscreen and number plate fitted.

*16)*Customer inspects vehicle and sign delivery note and collect documents (delivery note, sales deed, warranty certificate, horsepower, and insurance certificate).

V. Proposed System

Implementation of the vehicle importation, registration and purchase processes using the Blockchain technology can mitigate many of the challenges that were identified in the current system. The architecture diagram of Hyperledger Fabric vehicle server-side for the proposed system is shown in Fig. 4 in the appendix. The technologies used for the development of the proposed system are Spring Boot framework for server-side application and Angular 8+ for client-side application. Hyperledger Fabric vehicle application server consists of following tiers: (a) the web tier exposes web services for client application to consume. The client application sends a JSON request to the web tier which authenticates the request via a token management mechanism using OAuth2 and maps the JSON to java stub. The web layer communicates to the business layer which consists of several services (interfaces). (b) The business tier consists of 2 modules. These are the Hyperledger Fabric Vehicle Services which contains business requirements and the Hyperledger Fabric Vehicle Provider which is responsible for the IBM blockchain endpoints integration. Once the requested service is executed, the business tier sends the data to the web tier. The latter converts the object to JSON and sends to the client application. The Model View Component (MVC) framework has been adopted for the development of the client application. Each of these components is implemented to manage specific functionalities of the Hyperledger Vehicle web application. The description of the modules is as follows:

- AccountService is responsible for account functionalities; (a) login of user (b) registration of a participant (c) logout.
- VehicleCustomerSaleService is responsible for vehicle sale and registration flow.
- VehicleImportationService contains methods responsible for vehicle importation flow.
- VehicleSmartContractService invokes and queries transactions in the IBM Blockchain network.
- IbmBlockchainProvider is responsible for the IBM Blockchain web services integration.

The description (attributes) of the Smart Contract for Vehicle asset is as follows: VIN (vehicle identification number), make, model, type, color, year, engine number, engine capacity, mileage, fuel used, seat capacity, registration number, owner ID, owner name, owner address, amount, duty, description. The description of the methods of the smart contract is as follows:

- In this, method is executed when the smart contract is created.
- Invoke method is executed when a request is received to run a smart contract.
- CreateVehicle method creates the first block for a specific asset in the blockchain network.
- QueryAllVehicle method retrieves all the assets in the blockchain network.
- ChangeVehicleOwner method creates a new chain block with the owner's details.
- ChangeVehicleDetails method creates a new chain block with the modified vehicle's details.
- QueryVehicle method retrieves a vehicle transaction by the asset's key.
- GetHistoryOfVehicle method retrieves a historical chain of blocks for a specific vehicle.

The application interacts with the IBM Blockchain network to perform registration/enrollment, queries and updates. The smart contract contains functions that allow interacting with the ledger. Fig. 1 shows a representation of the application interaction with the ledger in the IBM blockchain network.

Enrolling the admin user: when the application is launched in the IBM blockchain network, an admin user is registered with the certificate authority. An enrollment call is triggered to the CA (Certificate Authority) server to retrieve the certificate (eCert) for this user. This enrollment certificate is used subsequently to register and enroll a new user.

Register and enroll a new user**:** a new user such as a manufacturer is created using the generated admin eCert. The latter is used to communicate with the CA server. This user <manufacturer> will be referenced as the identity to be used when querying and updating the ledger. It is important to highlight that the admin user's identity is used to issue the registration and enrollment calls for the new user. In other words, the user <manufacturer> is acting in the role of a registrar.

Querying the ledger: the data is stored as a series of key-value pairs. The application queries the ledger for the value of a single key or multiple keys. The data is returned in JSON format. Fig. 2 shows a representation of how a query works in the IBM Blockchain network.

Updating the ledger: The application creates a vehicle asset by the following steps: propose, endorse and notify the application. An order transaction is sent to the orderer and written to every peer's ledger. Fig. 3 shows the flow for updating a ledger on the IBM Blockchain network.

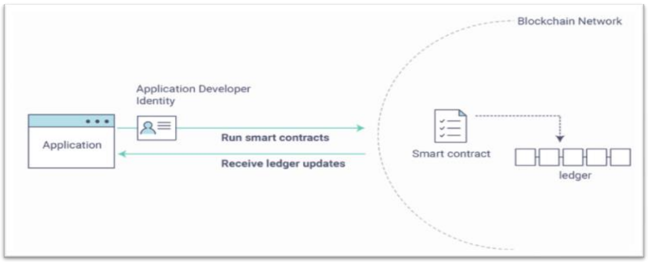


Fig. 1. Interactions in the IBM Blockchain Network.

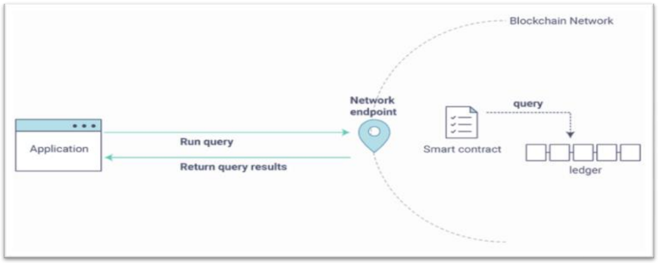


Fig. 2. Querying the Blockchain.

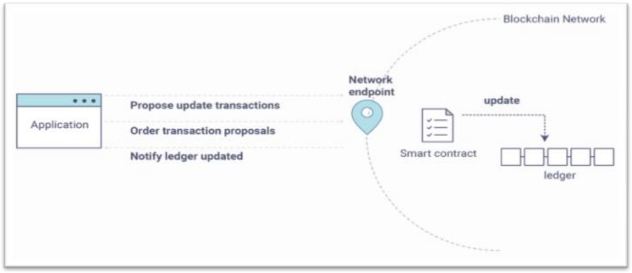


Fig. 3. Updating a Ledger on the Blockchain.

VI. Implementation of the Proposed System

The Hyperledger fabric vehicle application provides authentication mechanism for participants/peers in the IBM blockchain network. These participants are the manufacturer, auction, shipment, Mauritius Revenue Authority, car dealer, insurance, bank, registrar, National Transport Authority. Each participant has his/her own login credential to the system. Each participant has his/her assigned roles and responsibilities in the system. The access rights on the functionalities of the system are defined by the role of the participant. Upon successful authentication, the dashboard web page is loaded as shown in Fig. 5 (see appendix). It provides an overview of the system with the statistics of vehicle importations and sales and real-time information about the status and participants of the IBM Blockchain network.

A toolbox is always on display to ease the use for adding new process of vehicle importation and sale. The button <ADD VEHICLE IMPORTATION PROCESS> creates a new process flow is created in the database. The vehicle importation process provides a flow of the different stages and conditions involved when importing a vehicle and also the different participants/peers who are responsible in each stage. Moreover, the button <ADD VEHICLE SALES/REGISTATION PROCESS> creates a new process flow in the database for vehicle sales and registration after filling a form as shown in Fig. 6 (see appendix). The user (car dealer or admin) needs to provide information such as payment method and duty type. Likewise, the vehicle sales and registration process provide a flow of the different stages and conditions involved when a customer is proceeding with the deed of contract for the vehicle. Each stage is managed by different participants.

The import web page shows a list of vehicle importation processes. The different stages in the importation flow are tabulated with descriptions, rules, respective participants and status. For example, a car dealer needs to be complied with a set of rules prior to import vehicle in showroom. To create a vehicle record on the IBM Blockchain network, the user clicks on button <Go to Chain Code> and a form is loaded as shown in Fig. 7. The vehicle importation process consists of the following steps (Fig. 7 to 12 in appendix):

- The admin or manufacturer creates the genesis block in the IBM Blockchain network as shown in Fig. 7.
- Next, the admin or car dealer puts an offer and purchases the vehicle from the auction. The transaction is recorded into the blockchain as shown in Fig. 8.
- The asset needs to be shipped to the car dealer's physical address. Another block is created and linked to the chain into the IBM blockchain network as shown in Fig. 9.
- Once the asset is shipped to Mauritius, the asset is under Mauritius Revenue Authority's custody (bond). The car dealer needs to pay a duty to release the bond as shown in Fig. 10.
- Ownership of asset is transferred to the car dealer and the transaction is recorded on the IBM Blockchain network linked to the specific asset ID as shown in Fig. 11.
- Finally, the vehicle importation process is completed and is immutable as shown in Fig. 12.

The vehicle sale and registration dashboard web page provide a list of processes related to the vehicle sales as shown in Fig. 13 (see appendix). In order to sell a vehicle/asset to a customer, the user (car dealer or admin) clicks on button <Go to Chain Code> and fills the form as shown in Fig. 14 and Fig. 15 (see appendix). The vehicle sale and registration processed involve the following steps:

- The car dealer records the payment of the customer in the IBM Blockchain network as shown in Fig. 14.
- The car dealer creates a new block in the IBM blockchain network to transfer the asset's ownership to the customer as illustrated in Fig. 15.

The list of assets' transactions recorded in the IBM Blockchain network is provided on the Vehicle Asset dashboard web page as shown in Fig. 16 (see appendix). Each asset has its own historical records in the blockchain network which can be viewed when the user clicks on <VIEW HISTORY> as illustrated in Fig. 17 (see appendix).

The assessment of the POC blockchain-base smart contract application for automotive industry has been conducted as follows: (a) using the Hyperledger Fabric Vehicle application, we have demonstrated that the transaction histories of the vehicle are more transparent through the use of IBM blockchain technology where distributed ledger is shared among all network participants, (b) the Hyperledger Fabric Vehicle application provides an encryption mechanism to approve transactions. Falsification of mileage and other frauds are mitigated through the adoption of the IBM blockchain technology for the automotive industry. (c) We demonstrated that the purchase of vehicle needs to undergo a chain of processes and it can be very expensive and cumbersome to trace an item from its origin. Hence, through the application, it has become easy to audit trail the journey of an asset. Moreover, the application provides historical data that can be used to verify transactions. (d) Moreover, the Hyperledger Fabric vehicle application offers the ability for transactions or processes to be more efficiently through streamlining and automating. As demonstrated, the Hyperledger Fabric vehicle application provides a platform to make payments to different parties. The trust is gained from the data on the blockchains as the ledger is secure, error-free, accurate and immutable.

VII. Conclusion

Blockchain technology is still in a development phase and there are still discussions about its scalability. In this work, a Hyperledger Fabric Vehicle application has been developed that could solve the challenges and risks that the automotive industry is currently facing in the Republic of Mauritius. Due to the flexibility and modular consensus algorithm, the Hyperledger Fabric Vehicle can be further enhanced to implement other transactions for the automotive industry. The analysis of the application underlined that with the adoption of the IBM blockchain platform, we were able to achieve transparency, improve traceability, enhanced security, reduce costs and increase efficiency in the traditional system. Innovators have just started to scratch the surface of blockchain applications in the automotive industry of this long journey of disruptions.

References

[1] G. Ciatto, S. Mariani, A.Maffi and A. Omicini, “Blockchain-based Coordination: Assessing the Expressive Power of Smart Contracts”, Information, vol. 11, no. 52, pp. 1-20, January 2020. doi:10.3390/info11010052.

[2] S. Underwood, “Blockchain beyond Bitcoin”, Communications of the ACM, vol. 59, no. 11, pp. 15–17, October 2016. https://doi.org/10.1145/2994581.

[3] N. Szabo, “The idea of smart contracts. Nick Szabo’s Papers and Concise Tutorials”, 1997. Accessed on: Jan. 10, 2020. [Online]. Available: http://www.fon.hum.uva.nl/rob/Courses/InformationInSpeech

[4] QuillHash Team, “Looking Ahead: The Future of Automotive Industry and Blockchain Technology”, 2019. Accessed on: November 21, 2019. [Online]. Available: https://medium.com/quillhash.

[5] Défi Media Group, “Car Traffic: The Falsified Odometers for New Cars”, 2012. Accessed on: October 14, 2019. [Online]. Available: https://motors.mega.mu/news/car-traffic-falsified-odometers-new-cars-20120417.html.

[6] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System", 2008. Accessed on: October 28, 2019. [Online]. Available: https://bitcoin.org/bitcoin.pdf.

[7] Z. Zheng, S. Xie, H. Dai, X. Chen and H. Wang, “An Overview of Blockchain Technology: Architecture, Consensus and Future Trends”, in Proceedings of the 6th IEEE International Congress on Big Data, 25-30 June 2017, Honolulu, USA. doi: 10.1109/BigDataCongress.2017.85.

[8] S. Brakeville and B. Perepa, “Blockchain Basics: Introduction to Distributed Ledgers”, June 2019. Accessed on: Aug. 30, 2019. [Online]. Available: https://developer.ibm.com/technologies/blockchain/tutorials.

[9] X. Xu, C. Pautasso, L. Zhu, V. Gramoli, A. Ponomarev, A. B. Tran and S. Chen, "The Blockchain as a Software Connector", in Proceedings of the 13th IEEE/IFIP Conference on Software Architecture, 5-8 April 2016, pp. 182-191, Venice, Italy. doi: 10.1109/WICSA.2016.21.

[10] J. Stark, “Making sense of blockchain smart contracts”, June 4, 2016. Accessed on: September 30, 2019. [Online]. Available: http://www.coindesk.com/making-sense-smart-contracts/.

[11] K. Delmolino, M. Arnett, A. Kosba, A. Miller, and E. Shi, “Step by step towards creating a safe smart contract: Lessons and insights from a cryptocurrency lab," in Proceedings of the International Conference on Financial Cryptography and Data Security, 22-26 February 2016, pp. 79-94, Christ Church, Barbados.

[12] V. Morabito, “Business Innovation Through Blockchain: The B^3 Perspective”, 1st ed., Springer, 2017.

[13] D. Vujicic, d. Jagodic and S. Randic, “Blockchain technology, bitcoin, and Ethereum: A brief overview”, in Proceedings of the 17th IEEE International Symposium INFOTEH, 21-23 March 2018, East Sarajevo, Bosnia-Herzegovina. doi: 10.1109/INFOTEH.2018.8345547.

[14] N. Atzei, M. Bartoletti and T. Cimoli, "A survey of attacks on Ethereum smart contracts", in Proceedings of the 6th International Conference on Principles of Security and Trust, vol. 10204, pp. 164-186, 22-29 April 2017, Uppsala, Sweden.

[15] SecureLyte, “Introduction to Hyperlogic Fabric”, 2020. Accessed on: January 10, 2020. [Online]. Available: https://www.securelyte.com/introduction-to-hyperlogic-fabric/.

[16] SVR Technologies, “Blockchain Interview Questions and Answers”, 2020. Accessed on: January 10, 2020. [Online]. Available: https://svrtechnologies.com/blockchain-interview-questions-and-answers-pdf.

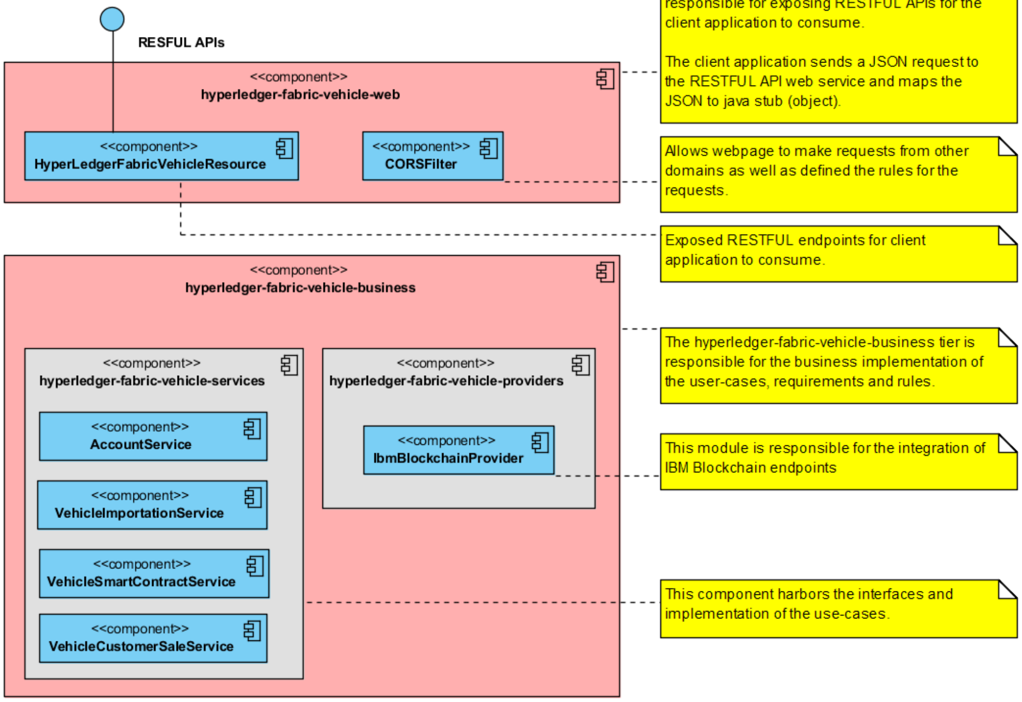
[17] R. Wolfson, “Ford Motor Company Launches Blockchain Pilot On IBM Platform to Ensure Ethical Sourcing of Cobalt”, January 16, 2019. Accessed on October 17, 2019. [Online]. Available: https://www.forbes.com/sites/rachelwolfson/2019/01/16/ford-motor-company-launches-blockchain-pilot-on-ibm-platform-to-ensure-ethical-sourcing-of-cobalt/#520837e75a1d.

[18] Business Blockchain HQ, “Automotive Blockchain News”, 2019. Accessed on October 17, 2019. [Online]. Available: https://businessblockchainhq.com/automotive-blockchain-news/.

[19] C. Torres, “Hyundai Subsidiaries Partner IBM to Accelerate AI and Blockchain Development”, February 20, 2019. Accessed on October 17, 2019. [Online]. Available: https://thecryptosight.com/hyundai-subsidiaries-partner-ibm-to-accelerate-ai-and-blockchain-development/.

[20] L. Mullan, “Hyundai Card and Hyundai Commercial partner with IBM to drive digital transformation”, February 21, 2019. Accessed on October 17, 2019. [Online]. Available: https://www.gigabitmagazine.com/ai/hyundai-card-and-hyundai-commercial-partner-ibm-drive-digital-transformation.

APPENDIX


RESFUL APIs
<<component>>
hyperledger-fabric-vehicle-web
<<component>>
HyperLedgerFabricVehicleResource
<<component>>
CORSFilter
<<component>>
hyperledger-fabric-vehicle-business
<<component>>
hyperledger-fabric-vehicle-services
<<component>>
AccountService
<<component>>
VehicleImportationService
<<component>>
VehicleSmartContractService
<<component>>
VehicleCustomerSaleService
<<component>>
hyperledger-fabric-vehicle-providers
<<component>>
IbmBlockchainProvider

The hyperledger-fabric-vehicle-web tier is responsible for exposing RESTFUL APIs for the client application to consume.

The client application sends a JSON request to the RESTFUL API web service and maps the JSON to java stub (object).

Allows webpage to make requests from other domains as well as defined the rules for the requests.

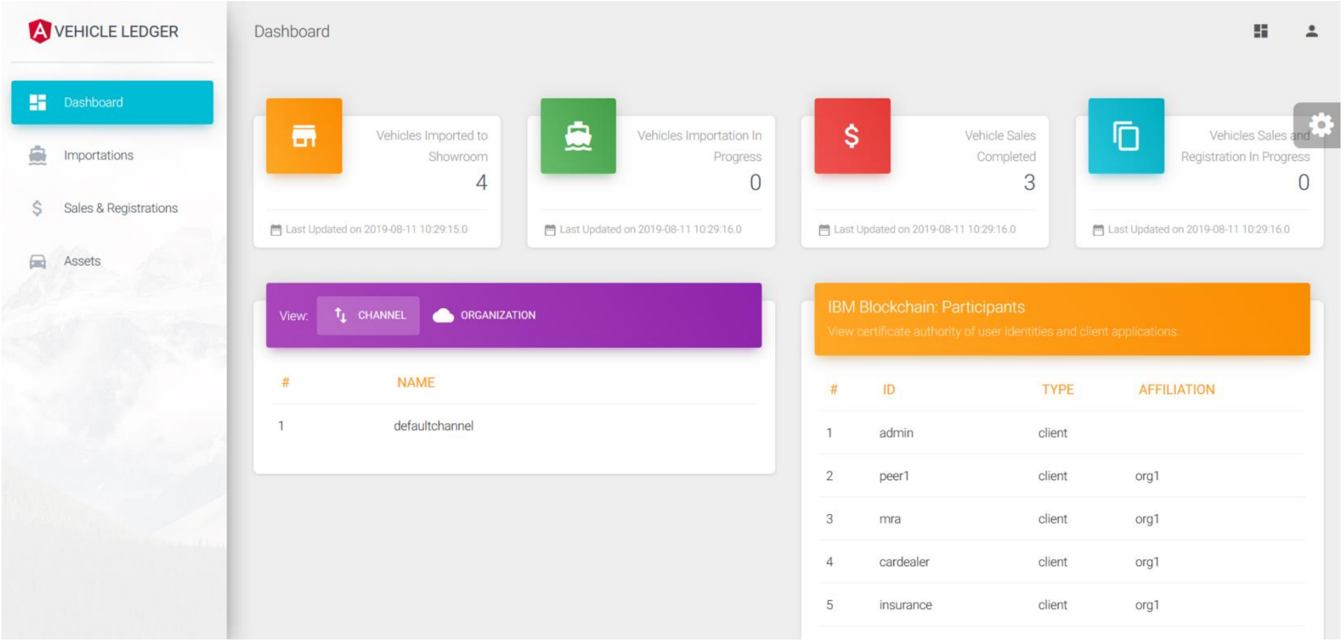
Exposed RESTFUL endpoints for client application to consume.

The hyperledger-fabric-vehicle-business tier is responsible for the business implementation of the user-cases, requirements and rules.

This module is responsible for the integration of IBM Blockchain endpoints

This component harbors the interfaces and implementation of the use-cases.

Fig. 4. Architecture Diagram for Server-Side Application for the Proposed System



Vehicles Imported to Showroom
4
Last Updated on 2019-08-11 10:29:15.0

Vehicles Importation In Progress
0
Last Updated on 2019-08-11 10:29:16.0

Vehicle Sales Completed
3
Last Updated on 2019-08-11 10:29:16.0

Vehicles Sales and Registration In Progress
0
Last Updated on 2019-08-11 10:29:16.0

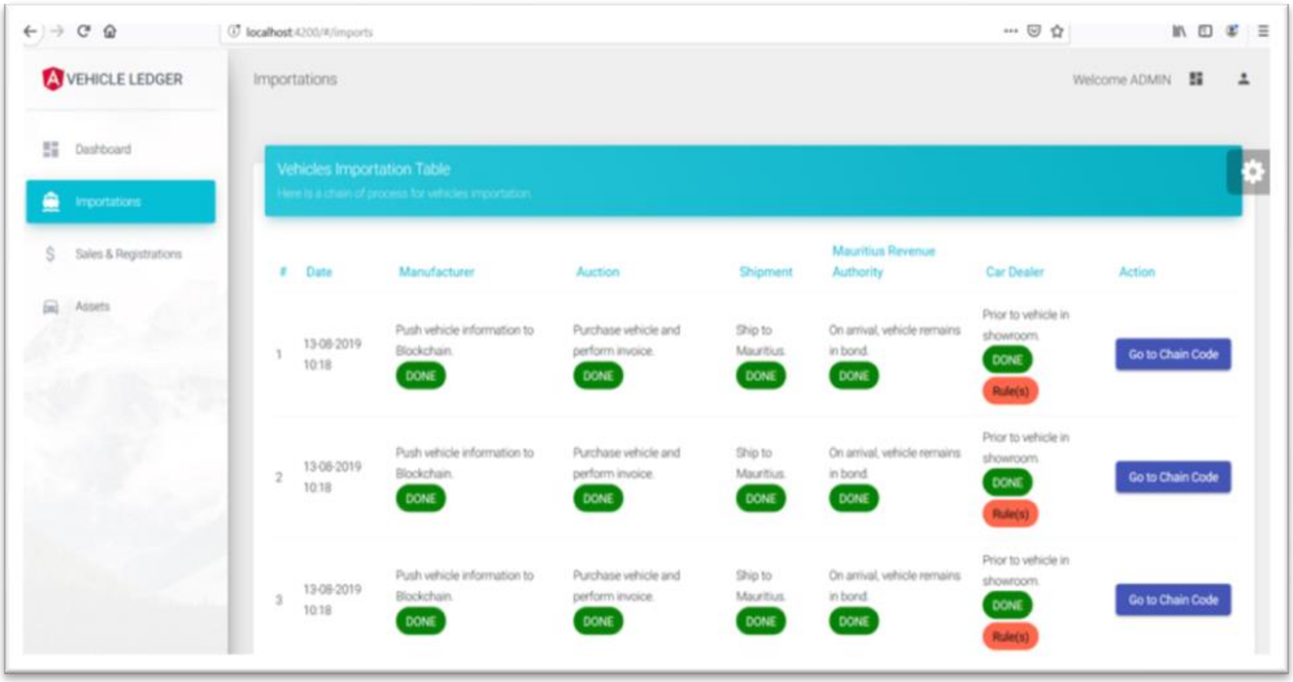
View: CHANNEL ORGANIZATION

#	NAME
1	defaultchannel

IBM Blockchain: Participants
View certificate authority of user identities and client applications

#	ID	TYPE	AFFILIATION
1	admin	client	
2	peer1	client	org1
3	mra	client	org1
4	cardealer	client	org1
5	insurance	client	org1

Fig. 5. Overview and Statistics of System.

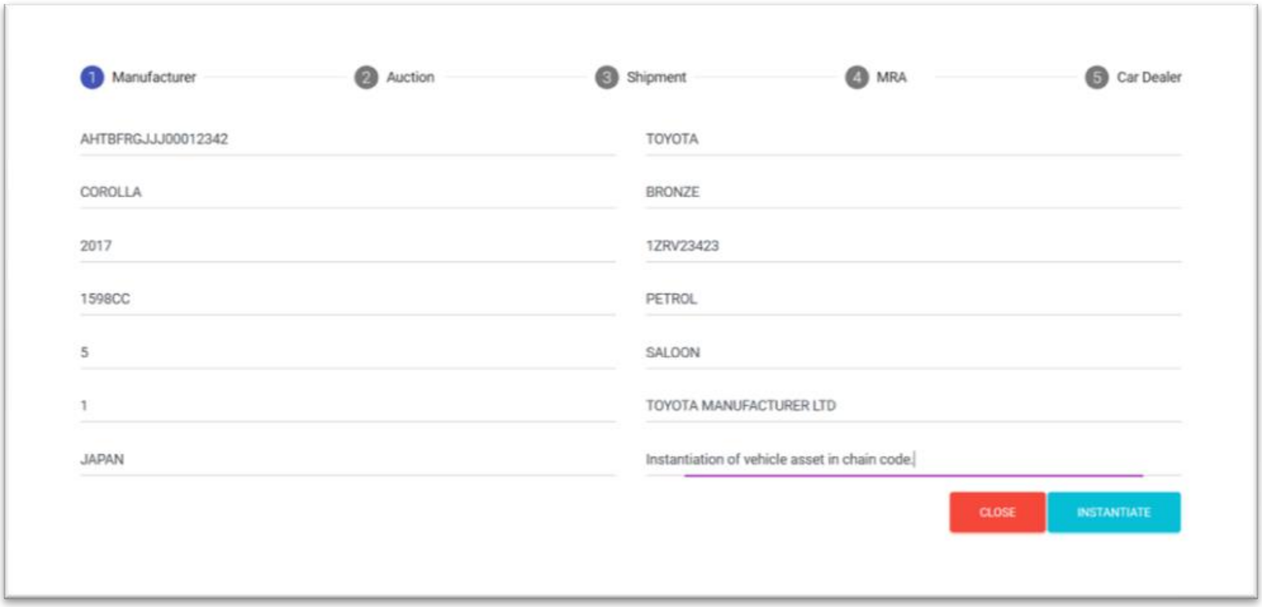


Vehicles Importation Table

Here is a chain of process for vehicles importation

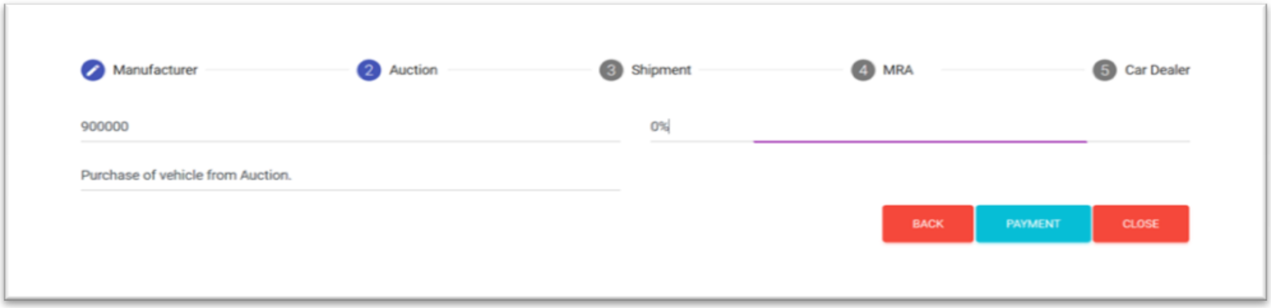
#	Date	Manufacturer	Auction	Shipment	Mauritius Revenue Authority	Car Dealer	Action
1	13-08-2019 10:18	Push vehicle information to Blockchain. DONE	Purchase vehicle and perform invoice. DONE	Ship to Mauritius. DONE	On arrival, vehicle remains in bond. DONE	Prior to vehicle in showroom. DONE Rule(s)	Go to Chain Code
2	13-08-2019 10:18	Push vehicle information to Blockchain. DONE	Purchase vehicle and perform invoice. DONE	Ship to Mauritius. DONE	On arrival, vehicle remains in bond. DONE	Prior to vehicle in showroom. DONE Rule(s)	Go to Chain Code
3	13-08-2019 10:18	Push vehicle information to Blockchain. DONE	Purchase vehicle and perform invoice. DONE	Ship to Mauritius. DONE	On arrival, vehicle remains in bond. DONE	Prior to vehicle in showroom. DONE Rule(s)	Go to Chain Code

Fig. 6. Vehicle Import Page.



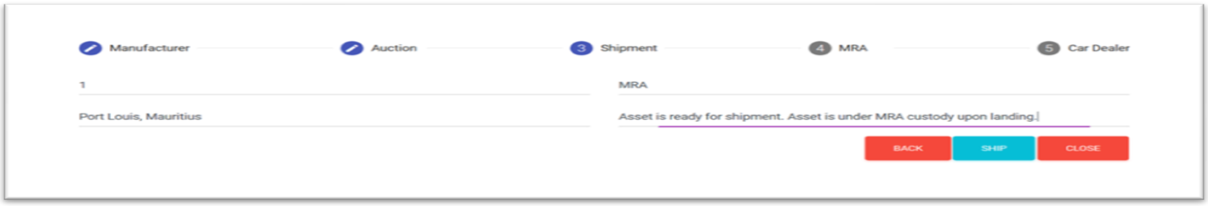
AHTBFRGJJJ00012342	TOYOTA
COROLLA	BRONZE
2017	1ZRV23423
1598CC	PETROL
5	SALOON
1	TOYOTA MANUFACTURER LTD
JAPAN	Instantiation of vehicle asset in chain code.

Fig. 7. Vehicle Importation - Manufacturer Creates Genesis Block in Blockchain.



900000	0%
Purchase of vehicle from Auction.	

Fig. 8. Vehicle Importation – Purchase Asset from Auction.



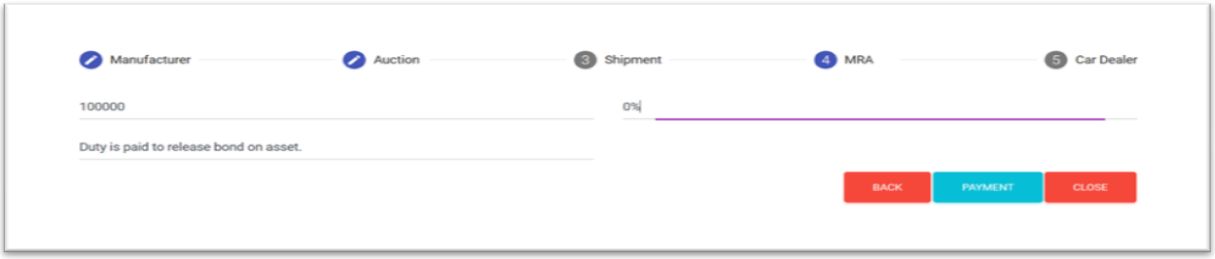
1

Port Louis, Mauritius

MRA

Asset is ready for shipment. Asset is under MRA custody upon landing.

Fig. 9. Vehicle Importation – Asset Ready for Shipment.

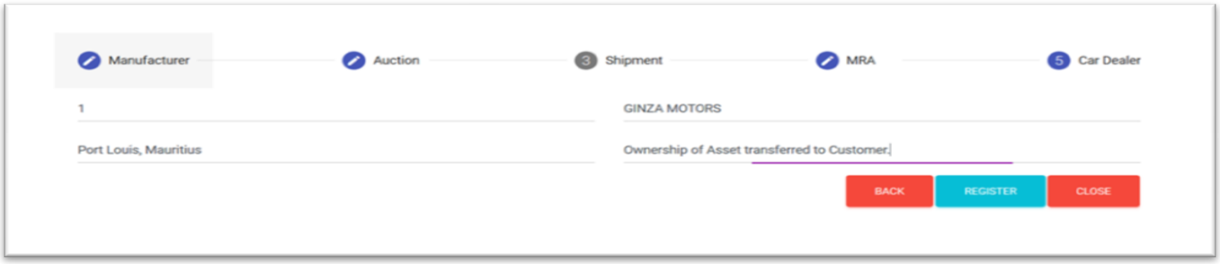


100000

Duty is paid to release bond on asset.

0%

Fig. 10. Vehicle Importation – Payment to Release Custody from MRA.



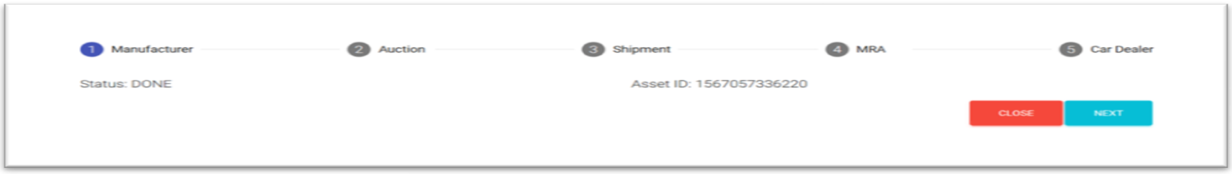
1

Port Louis, Mauritius

GINZA MOTORS

Ownership of Asset transferred to Customer.

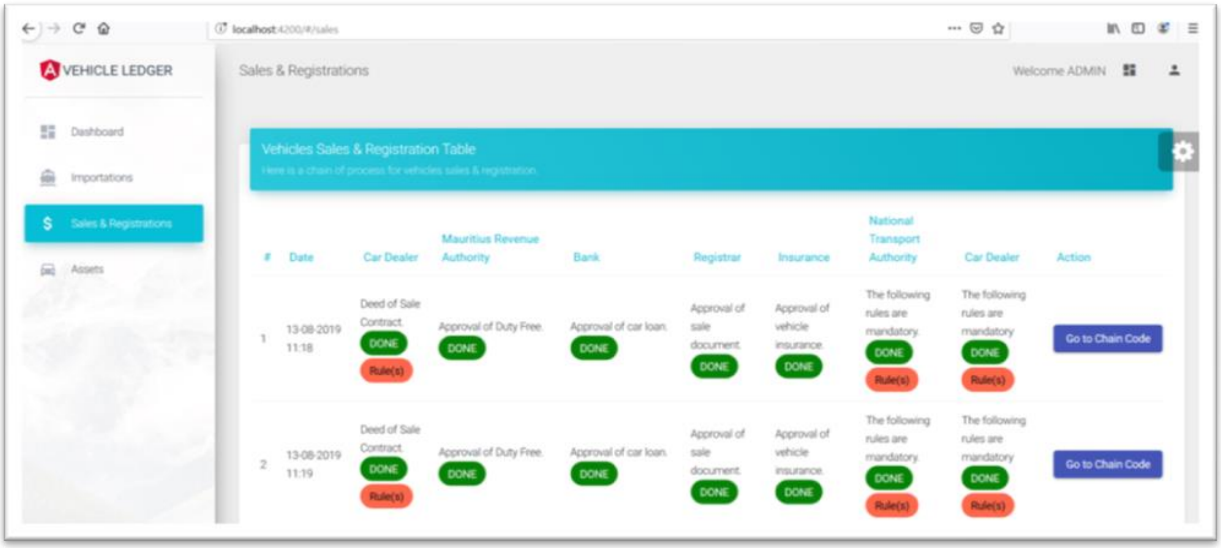
Fig. 11. Vehicle Importation – Transfer Ownership of Asset to Car Dealer.



Status: DONE

Asset ID: 1567057336220

Fig. 12. Vehicle Importation Completed and Immutable.

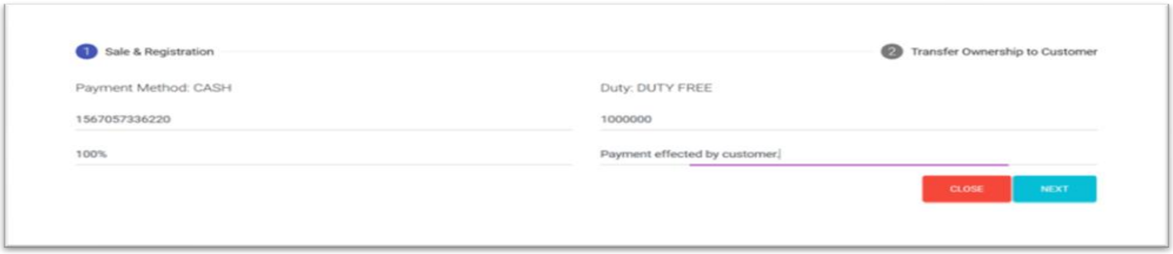


Vehicles Sales & Registration Table

Here is a chain of process for vehicles sales & registration

#	Date	Car Dealer	Mauritius Revenue Authority	Bank	Registrar	Insurance	National Transport Authority	Car Dealer	Action
1	13-08-2019 11:18	Deed of Sale Contract. DONE Rule(s)	Approval of Duty Free. DONE	Approval of car loan. DONE	Approval of sale document. DONE	Approval of vehicle insurance. DONE	The following rules are mandatory. DONE Rule(s)	The following rules are mandatory. DONE Rule(s)	Go to Chain Code
2	13-08-2019 11:19	Deed of Sale Contract. DONE Rule(s)	Approval of Duty Free. DONE	Approval of car loan. DONE	Approval of sale document. DONE	Approval of vehicle insurance. DONE	The following rules are mandatory. DONE Rule(s)	The following rules are mandatory. DONE Rule(s)	Go to Chain Code

Fig. 13. Vehicle Sale and Registration Dashboard Page.



Payment Method: CASH

Duty: DUTY FREE

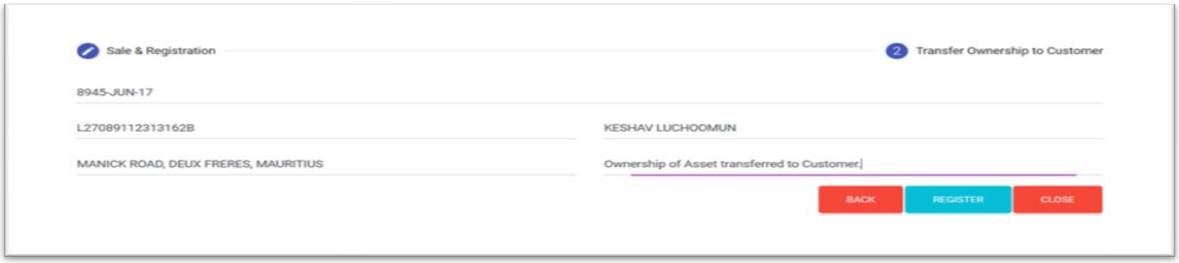
1567057336220

1000000

100%

Payment effected by customer.

Fig. 14. Vehicle Sale – Payment by Customer.



8945-JUN-17

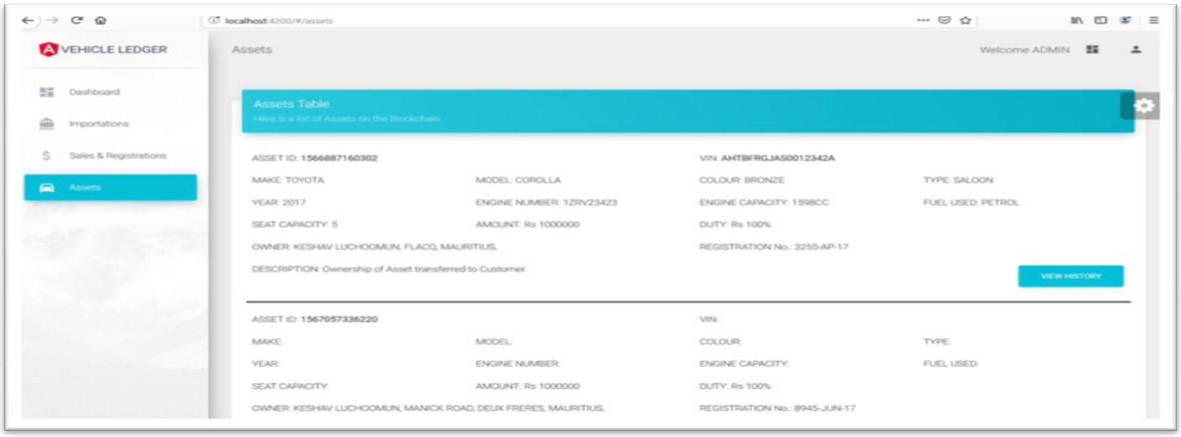
L27089112313162B

KESHAV LUCHOOMUN

MANICK ROAD, DEUX FRERES, MAURITIUS

Ownership of Asset transferred to Customer.

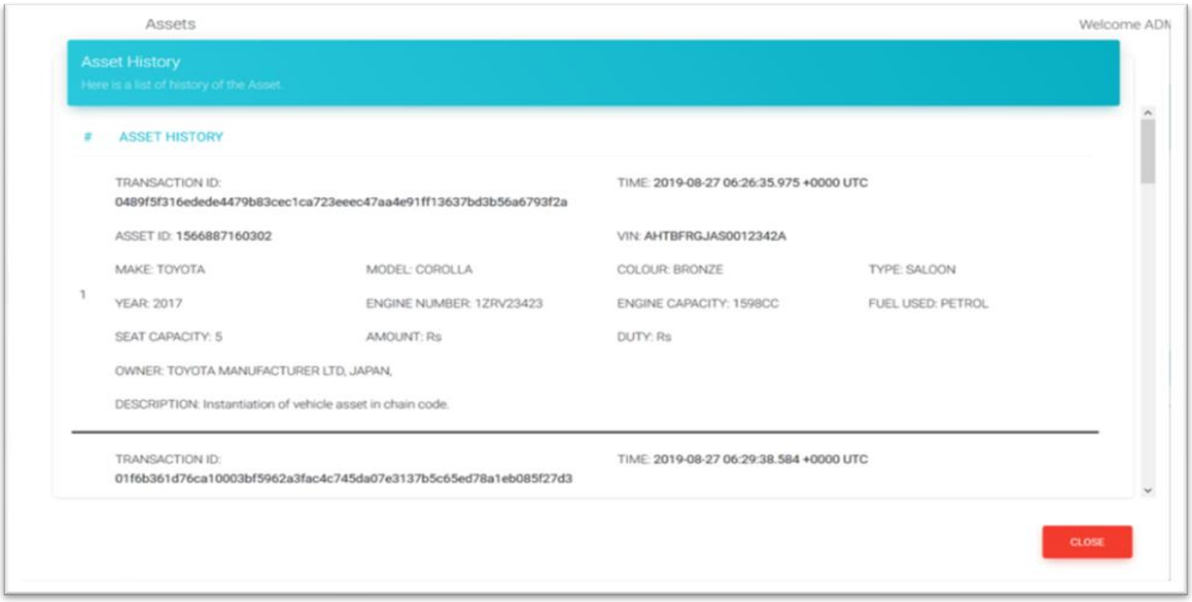
Fig. 15. Vehicle Sale – Transfer of Vehicle's Ownership to Customer.



ASSET ID: 1566887160302
VIN: AHTBFRGJAS0012342A
MAKE: TOYOTA
MODEL: COROLLA
COLOUR: BRONZE
TYPE: SALOON
YEAR: 2017
ENGINE NUMBER: 1ZRV23423
ENGINE CAPACITY: 1598CC
FUEL USED: PETROL
SEAT CAPACITY: 5
AMOUNT: Rs 1000000
DUTY: Rs 100%
OWNER: KESHAV LUCHOOMUN, FLACQ, MAURITIUS,
REGISTRATION No.: 3255-AP-17
DESCRIPTION: Ownership of Asset transferred to Customer.

ASSET ID: 1567057336220
VIN:
MAKE:
MODEL:
COLOUR:
TYPE:
YEAR:
ENGINE NUMBER:
ENGINE CAPACITY:
FUEL USED:
SEAT CAPACITY:
AMOUNT: Rs 1000000
DUTY: Rs 100%
OWNER: KESHAV LUCHOOMUN, MANICK ROAD, DEUX FRERES, MAURITIUS,
REGISTRATION No.: 8945-JUN-17

Fig. 16. Vehicle Asset – List of Assets in the IBM Blockchain Network.



1

TRANSACTION ID:
0489f5f316ededе4479b83cec1ca723eeec47aa4e91ff13637bd3b56a6793f2a
TIME: 2019-08-27 06:26:35.975 +0000 UTC
ASSET ID: 1566887160302
VIN: AHTBFRGJAS0012342A
MAKE: TOYOTA
MODEL: COROLLA
COLOUR: BRONZE
TYPE: SALOON
YEAR: 2017
ENGINE NUMBER: 1ZRV23423
ENGINE CAPACITY: 1598CC
FUEL USED: PETROL
SEAT CAPACITY: 5
AMOUNT: Rs
DUTY: Rs
OWNER: TOYOTA MANUFACTURER LTD, JAPAN,
DESCRIPTION: Instantiation of vehicle asset in chain code.

TRANSACTION ID:
01f6b361d76ca10003bf5962a3fac4c745da07e3137b5c65ed78a1eb085f27d3
TIME: 2019-08-27 06:29:38.584 +0000 UTC

Fig. 17. Vehicle Asset – view Historical Records of a Specific Asset.

Method for Rainfall Rate Estimation with Satellite based Microwave Radiometer Data

Kohei Arai
Graduate School of Science and Engineering
Saga University, Saga City
Japan

***Abstract*—Method for rainfall rate estimation with satellite based microwave radiometer data is proposed. A method to consider the geometric relationship of the observed ice particles and microwave radiometer in the estimation of precipitation is shown, and its validity is shown by comparing it with precipitation radar data on the ground. Observations at high altitudes, such as ice particles, differ greatly in the location of the observation point projected on the ground surface and in the upper troposphere where the observations exist. This effect was insignificant when the precipitation was small because ice particles were often absent, but it was found that the effect was large when the precipitation was large. In other words, the proposed method is effective and effective for Advanced Microwave Scanning Radiometer (AMSR) data in Houston, which was shown as an example of a highly developed convective rain cloud with an In the case of Kwajalein, the effect is insignificant. In addition, the proposed method requires an assumption of ice particle height, and it is necessary to make assumptions based on climatic values. In addition, microwaves in the 89 GHz band, which are considered to be sensitive to ice particles, are not only sensitive to ice particles, so it must be taken into account that they are also affected by the presence of non-ice particles.**

Keywords*—*Rainfall rate estimation; Advanced Microwave Scanning Radiometer: AMSR; geometric relation

I. Introduction

Spencer et al., Liu, and Curry proposed a precipitation estimation algorithm using a microwave radiometer Special Sensor Microwave / Imager (SSM / I) [1]. In addition, Liu modified the algorithms such as the beam filling effect (Beam filling effect) to these algorithms, and the microwave radiometer Earth Observation Satellite System: EOS: AMSR-E mounted on the Earth observation satellite Advanced Earth Observing Satellite: ADEOS Ⅱ has the same sensor as the AMSR itself [2]. Therefore, AMSR and AMSR-E can use the same estimation algorithm.

Current precipitation of AMSR-E The Japan Aeronautics Exploration Agency: JAXA standard product uses a precipitation estimation algorithm by Liu [3]. In precipitation estimation using microwave radiation, the observed brightness temperature in the frequency band around 18 GHz and 89 GHz is generally used [4]. That is, the absorption attenuation of microwaves in the frequency band around 18 GHz and the scattering by ice particles in the frequency band around 89 GHz are taken into account [5].

Considering the vertical structure of a rain cloud accompanied by strong precipitation, such a rain cloud has a liquid layer up to 0 ° altitude (freezing altitude) at the bottom, and a mixture of solid (ice particles) and liquid (rain). It consists of a layer and the top layer consisting only of solids (ice particles). Compared to stratified clouds, convective clouds with strong updrafts generally produce more ice particles due to the updrafts, cause strong rainfall, and have large spatiotemporal changes [6].

The brightness temperature of the 89 GHz band microwave is dominated by the scattering of ice particles in the mixed layer and the top layer, and the brightness temperature of the 18 GHz band microwave is dominated by the absorption and attenuation due to the raindrops in the lower layer. Therefore, the microwaves in the 89 GHz band and the 18 GHz band have different altitudes of sensitivity to precipitation [7].

The microwave radiometer AMSR-E on board the Aqua satellite (the second satellite of EOS satellite series) has a footprint in the 18 GHz band of 16 × 27 km (scanning × traveling direction), which is too coarse to observe local rain due to convective clouds such as cumulonimbus clouds . For this reason, local precipitation estimation was attempted using a high-frequency band with a relatively small footprint. However, as described above, ice particles sensitive to the high-frequency band exist at high altitudes. It is necessary to make it consistent with the surface precipitation.

In this study, the author attempted to estimate precipitation from the effect of scattering from ice particles in clouds using 89 GHz data of AMSR-E. As a result, it was found that there was a displacement between the estimated precipitation and the distribution of precipitation measured by ground-based precipitation radar.

The next section, related research works are reviewed followed by proposed method with theoretical background. Then, some experimental works are described together with some results. After that, conclusion and some remarks are described together with future research works.

II. Related Research Works

Evaluation of Marine Observation Satellite: MOS-1 Microwave Scanning Radiometer: MSR data in field experiments was made for microwave radiometer performance evaluation [8]. Advanced Microwave Scanning Radiometer: AMSR is proposed and the preliminary study is made [9] together with specification and design of the instrument [10]. On the other hand, antenna pattern correction and Sea Surface Temperature: SST estimation algorithms for AMSR is proposed [11] together with potential capabilities of the instrument [12].

Meanwhile, simultaneous estimation of sea surface temperature, wind speed and water vapor with AMSR-E data based on improved simulated annealing is proposed [13]. Also, correction of the effect of relative wind direction on wind speed derived by AMSR is proposed and validated [14]. Precipitation estimation using AMSR data considering geometric relationship between observation target and radiometer is proposed [15]. Nonlinear optimization based SST estimation methods with remote sensing satellite based Microwave Scanning Radiometer: MSR data is proposed and validated [16].

As for rainfall rate estimation, method for estimation of rain rate with Rayleigh and Mie scattering assumptions on the Z-R relationship for different rainfall types is proposed and evaluated its accuracy [17]. On the other hand, comparison between Rayleigh and Mie scattering assumptions for Z-R relation and rainfall rate estimation with Tropical Rainfall Measuring Mission: TRMM/Precipitation Radar: PR data is proposed [18]. Detecting algorithm for rainfall area movement based on Kalman Filtering is also proposed [19]. Meanwhile, reconstruction of cross section of rainfall situations with precipitation radar data based on wavelet analysis is proposed and well reported [20].

III. Proposed Method

The JAXA standard algorithm described above, which is used for AMSR-E standard products, is based on the Liu algorithm (hereinafter referred to as the standard algorithm). This standard algorithm does not take into account the altitude of the object to be observed or the orientation of the sensor mounted on the satellite. There is a report that clarifies the effect of the radiative transfer equation on the upward luminance temperature from the surroundings on the footprint of interest, taking into account the incident angle of 53 degrees, such as SSM / I. Some reports have evaluated the effects of slant paths on microwave radiometers.

A. Observation Configuration

In this study, the author propose a method to consider the observation configuration in the estimation of precipitation. In other words, the shift caused by observation using microwaves in the 89 GHz band was attributed to the altitude of the target and the direction of observation, and the author proposed precipitation estimation taking this into account. Fig. 1 shows geometric relation between AMSR and observation target of clouds.

Emitted electronic and magnetic wave from the observation area is propagated with incidence angle through a rainy atmosphere including ice particles in clouds and raindrops reached to the microwave radiometer onboard satellite. In this observation configuration, there is a displacement between actual observation area and apparent observation area. Therefore, this displacement causes same estimation errors in precipitation and rainfall rate estimations.

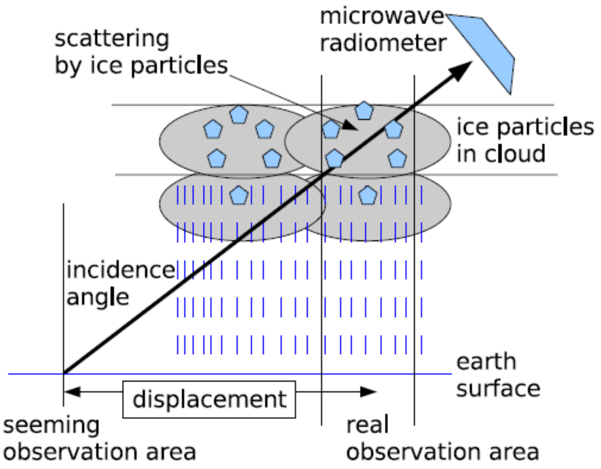


Fig. 1. Geometric Relation between AMSR and Observation Target of Clouds.

Assuming that the difference between the observation points due to the observation configuration depends on the average altitude at which ice particles exist (called the average height of ice particles), the average height of ice particles, the incident angle and azimuth of observation by the microwave radiometer, and the apparent Between the actual observation point and the difference between the observation point and the latitude / longitude is expressed as follows,

$$lat, = lat' + D\cos\theta_a$$
$$lon = lon' + D\sin\theta_a$$
$$D = H / \tan(90 - \theta i) \qquad (1)$$

where *lat, lon* is the latitude and longitude of the actual observation point, *lat, lon* is the latitude and longitude of the apparent observation point, *D* is the distance of horizontal deviation, *H* is the average ice particle height, *θa* is the observation azimuth, *θi* is the observed incident angle.

B. Conventional Method

Precipitation estimation algorithm by Liu (The relationship between precipitation *R* and microwave index *f*) was expressed using the beam-filling effect.

$$R = \alpha f^{\beta} \qquad (2)$$

α and *β* are called beam-filling factors and are factors that depend on the footprint size of the sensor. The footprint of the 19GHz band of SSM / I is less than 50km, *α* = 10.6 and *β* = 1.621. AMSR's 18GHz band footprint is about half that of SSM / I, and *α* = 8.25, *β* = 1.88, respectively. Microwave index *f* is represented as follows,

$$f = (1 - D/D_0) + 2(1 - PCT/PCT_0) \qquad (3)$$

$$D = T_{B18V} - T_{B18H}$$

$$PCT = 1.818\ T_{B5V} - 0.818\ T_{B5H} \qquad (4)$$

The first term on the right-hand side represents microwave radiation absorption by raindrops, and the second term represents the effect of microwave scattering by ice particles contained in rain clouds. *D* is the difference between vertical and horizontal polarization (Depolarization), *PCT* is the polarization-corrected temperature.

T_B is the luminance temperature observed by the sensor, the subscript number is frequency, the subscript V is vertical polarization, and the subscript H is horizontal polarization. D_0 and PCT_0 are thresholds at the beginning of rain, respectively, and a monthly reference table is created using the observed brightness temperature and sea surface temperature of 36.5 GHz.

D and *PCT* are compatible with SSM / I, when used in AMSR and AMSR-E as follows,

$$D_{SSM/I} = -0.14 + 0.903D_{AMSR}$$

$$PCT_{SSM/I} = 2.2 + 0.996\ PCT_{AMSR} \tag{5}$$

Since the beam filling factors α and β correspond to the 18 GHz observation data, the physical quantity obtained by this method is the instantaneous surface precipitation that is averaged within the footprint of the 18 GHz observation microwave. By the way, the footprint of the 18 GHz observation microwave is 15.7 × 27.4 [km] (scanning direction × traveling direction).

C. Proposed Method

In order to fix the beam filling factors α and β that are indicators of raindrops, we estimate precipitation using only the scattering of ice particles, taking into account the dynamic range of *f*.

The effects of raindrop absorption and ice particle scattering are used for microwave precipitation estimation, but by removing the raindrop absorption element (18 GHz) therefrom, precipitation that considers only ice particle scattering is taken into account. Precipitation can be estimated (precipitation using only the observed brightness temperature in the 89 GHz band). Then, equation (3) can be replaced to equation (6).

$$f = 4(1 - PCT/PCT_0) \tag{6}$$

Calculate the deviation within a certain area and the average height of ice particles based on the average height of ice particles and the correlation coefficient between actual precipitation and scattered precipitation of ice particles. The correlation coefficient is defined as follows,

$$R = \frac{\sum_N (P_t - P_{t'})(P_i - P_{i'})}{\sqrt{\sum_N (P_t - P_{t'})^2}\sqrt{\sum_N (P_i - P_{i'})^2}} \tag{7}$$

where *N* is the number of data, *Pt* is the actual precipitation, *Pi* is the scattered ice particle precipitation. Also, *Pt'* and *Pi'* are mean of *Pt'* and *Pi'*, respectively.

When the correlation reaches the peak, calculate the deviation within a certain area and the average height of ice particles. Also, precipitation radar observations and AMSR-E observations estimate gaps and ice particle average altitudes using spatiotemporal matchup data.

IV. Experiment

A. Data used

As the actual precipitation, the author estimate the surface precipitation obtained from ground precipitation radars and rain gauges. As the actual precipitation, the surface precipitation (L2A52, NASA TRMM Ground-based Instrument Data) obtained from ground precipitation radars and rain gauges is used. The radar sites are Houston, Texas (N29.472 [degree], W95.079 [degree]) as the subtropical zone and Kwajalein, N8.720 [degree], E167.740 [degree] as the tropical zone, Marshall Islands).

The precipitation from ground-based precipitation radar is referred to as "actual precipitation", and the estimated precipitation due to ice particle scattering is referred to as "ice particle scattering precipitation". In order to take into account the effects of seasonal fluctuations, etc., and to improve the reliability of the analysis results, the year in Houston in 2008 was assumed. (Houston in July and Kwajalein in September).

B. Area in Concern

Since the estimation accuracy of ice particle scattered precipitation is not very good, the difference from the actual precipitation is remarkable. However, the distribution of precipitation in a certain area of ice particle scattered precipitation and actual precipitation shows similar characteristics. By the optimization method using the distribution, it is possible to examine the difference between the actual precipitation and the ice particle scattering precipitation. The problem with this method is that it must be assumed that the shift within the constant area used is uniform. Therefore, it is assumed that the ice particle height within a certain area is also uniform. The smaller the constant area used here, the better, but if it is too small, it will not find the distribution.

In the precipitation estimation on January 10, 2008, the analysis was divided into four areas as shown in Fig. 2.

As a result, the estimated ice crystal flat altitudes showed different values as shown in Fig. 3. For this reason, it can be said that the smaller the fixed area used here is, the more preferable it is. However, if it is too small, it becomes impossible to find a distribution pattern. In addition, the probability that a cloud of the same pattern is generated increases, and it is likely that a local optimal solution is likely to occur. For these reasons, it is difficult to determine the optimal range, so in this study this fixed area was used as the measurement range for ground-based precipitation radar. Correlation coefficients for the designated four areas as a function of scattering altitude is shown in Fig. 3.

The fixed area used in this study is the measurement range of ground precipitation radar. In this study, it is assumed that the cause of the slip is the observation configuration. The actual cause of this shift is also caused by the parallel movement by the wind and the falling time of precipitation. In this study, we assumed the average height of ice particles and used it as a variable. The time when the correlation coefficient between the actual precipitation and the ice particle scattered precipitation is the highest is the time when the distribution is the best.

As an example, Fig. 4 shows the correlation coefficient when the average height of ice particles in the data of Houston on July 2, 2007 was changed.

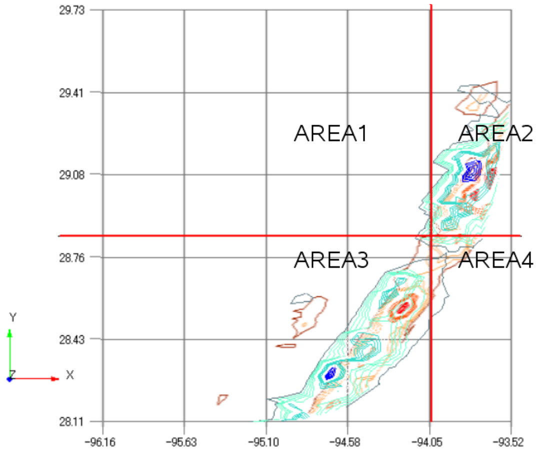


Fig. 2. Estimated Rainfall Rate with AMSR-E Data Acquired on January 10 2008 for Four Divided Areas.

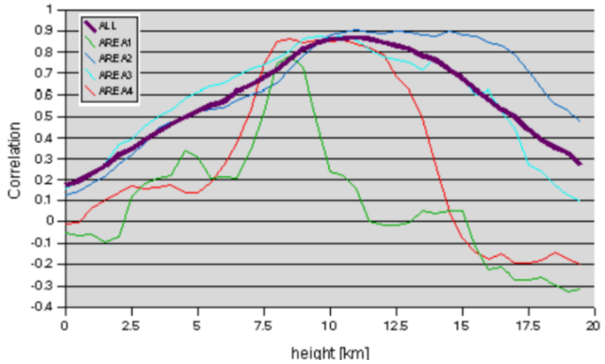


Fig. 3. Correlation Coefficients for the Designated Four Areas as a Function of Scattering Altitude.

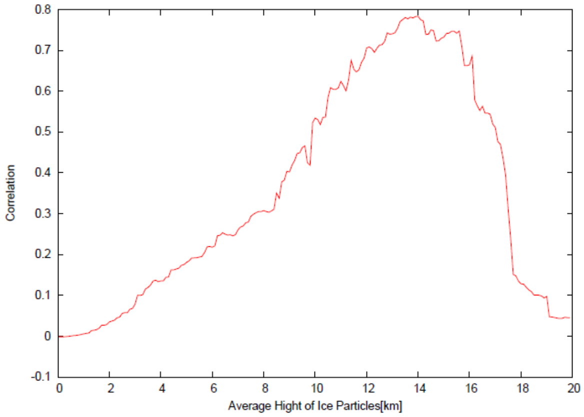


Fig. 4. Relation between Average Height of Ice Particles and Correlation between Actual and Estimated Precipitation for the Observation Data of July 2 2007.

It can be said that it is possible to estimate the altitude at which ice particles are present from the geometric relationship showing the peak of the correlation coefficient. Fig. 4 shows the correlation between the change in the average ice particle height and the actual precipitation distribution in the observation data on July 2, 2007. The horizontal axis is the assumed ice particle height and the vertical axis is the correlation coefficient between the actual precipitation and the ice particle scattering precipitation.

C. *Estimated Results*

In this study, the gap between precipitation radar observation and AMSR-E observation was estimated using spatiotemporal matchup data, and the average height of ice particles was estimated. Table I shows the results for Houston in July 2007, and Table II shows the results for Kwajalein in September 2007.

In the table, the number of data indicates the number of constant regions of AMSR-E for which precipitation was estimated, and Rt (mm / hr) is the average precipitation in the observation area. Also, Orig.R shows the correlation coefficient between the actual precipitation and the estimated precipitation when the height of the ice particles is not taken into account, and Max.R shows the correlation coefficient when taking that into account. H and D indicate the height of the ice particle level and the amount of deviation from the position showing the maximum correlation with the observation position. N / A indicates not applicable. In this case, the estimated precipitation is close to 0, which means that it is excluded from consideration.

TABLE. I. DISTANCE BETWEEN OBSERVED LOCATION AND THE LOCATION THAT SHOWS HIGHEST CORRELATION OF ESTIMATED PRECIPITATION, AND ESTIMATED MEAN ALTITUDE OF ICE PARTICLES (HOUSTON DATA)

Data No.	Date	No. of Data	Rt(mm/hr)	Orig R	Max R	H(km)	D(km)
1	7/2	676	0.24	0	0.78	14	0.18
2	7/4	627	9.03	0.34	0.64	10.5	0.15
3	7/5	599	0.51	0.06	0.23	3.5	0.04
4	7/6	741	0.8	0.06	0.66	13.5	0.17
5	7/7	1032	0	0.06	0.06	0	0
6	7/7	541	0	N/A	N/A	0	0
7	7/9	661	0	-0.01	-0.01	0	0
8	7/9	562	0	N/A	N/A	0	0
9	7/11	730	0	N/A	N/A	0	0
10	7/12	730	0	N/A	N/A	0	0
11	7/13	638	0	N/A	N/A	0	0
12	7/14	574	0	N/A	N/A	0	0
13	7/16	793	0.03	0.12	0.43	6	0.08
14	7/16	543	0	N/A	N/A	0	0
15	7/18	450	3.19	0.14	0.61	19	0.24
16	7/20	627	0.12	0.02	0.69	11	0.14
17	7/21	628	0	N/A	N/A	0	0
18	7/22	684	0	N/A	N/A	0	0
19	7/23	1091	0	0.09	0.63	12.5	0.16
20	7/23	543	0	N/A	N/A	0	0
21	7/25	676	0.13	0.15	0.51	8	0.1
22	7/25	471	0.75	0.05	0.47	11.5	0.12
23	7/27	684	2.33	0.16	0.68	17.5	0.22
24	7/27	712	0	0.11	0.21	0.5	0.18
25	7/28	720	0.68	0.13	0.21	2	0.03
26	7/30	568	0	N/A	N/A	0	0

TABLE. II. DISTANCE BETWEEN OBSERVED LOCATION AND THE LOCATION THAT SHOWS HIGHEST CORRELATION OF ESTIMATED PRECIPITATION, AND ESTIMATED MEAN ALTITUDE OF ICE PARTICLES (KWAJALEIN DATA)

Data No.	Date	No. of Data	Rt(mm/hr)	Orig R	Max R	H(km)	D(km)
1	9/1	931	0.75	0.28	0.51	11.6	16.26
2	9/1	939	0.29	0.33	0.56	8.6	12.06
3	9/3	909	0.51	0.22	0.43	9	12.62
4	9/3	1146	0.5	0.01	0.35	13	18.23
5	9/4	1070	0.14	0.06	0.16	11.2	15.7
6	9/5	902	0.22	0.04	0.39	14.8	20.75
7	9/5	174	0.29	0.04	0.76	13.4	18.79
8	9/6	404	0.2	0.05	0.06	2.6	3.65
9	9/6	1041	0.17	N/A	N/A	0	0
10	9/8	919	0.48	0.33	0.45	3	4.21
11	9/10	1002	0.55	0.35	0.62	9.2	12.9
12	9/12	897	0.24	0.17	0.22	17.2	24.11
13	9/12	956	0.16	0.02	0.02	2.8	3.93
14	9/13	1018	0.15	0.43	0.47	3.8	5.33
15	9/13	1231	0.17	0.12	0.13	7.4	10.37
16	9/15	1045	4.29	0.73	0.78	5.8	8.13
17	9/15	958	0.48	0.43	0.53	10.4	14.58
18	9/17	936	0.25	-0.01	0.06	7.2	10.09
19	9/19	1000	5.89	0.54	0.75	15.2	21.31
20	9/20	1071	2.18	0.73	0.78	12.2	17.1
21	9/21	992	0.14	0.07	0.33	9	12.62
22	9/22	1003	1.43	0.42	0.5	9	12.62
23	9/22	1042	0.16	0.16	0.19	9.4	13.18
24	9/24	966	0.2	0.26	0.29	3.4	4.77
25	9/26	945	0.3	0.16	0.28	13	18.23
26	9/26	723	0.71	0.34	0.6	11	15.42
27	9/28	1130	0.43	0.29	0.47	10	14.02
28	9/28	936	0.41	0.34	0.51	9.2	12.9
29	9/29	381	0.15	N/A	N/A	0	0
30	9/29	915	0.1	N/A	N/A	0	0
31	9/30	204	1.78	0.09	0.74	20	28.04

According to Tables I and II, data with high average precipitation in the observation area are assumed assuming the average height of ice particles. The number of relationships has improved significantly. This is because the distribution becomes clearer as the precipitation increases. Conversely, for data with low precipitation, there is no characteristic of the distribution of precipitation, so no improvement in the correlation coefficient is observed. This may be due to the absence of ice particles in rain clouds when precipitation is low.

The details of the data on July 2, 2007 in Houston are shown below as examples showing the distribution and deviation of each data and the ice particle level and altitude. First, Fig. 5 shows the distribution of actual precipitation and ice particle scattering precipitation. From Fig. 5, it can be seen that the distribution of actual precipitation and the scattering of ice particle precipitation roughly agree, but a displacement of about 10 to 20 km occurs. Observation in the 89 GHz band has a small footprint (5.9 km in the direction of observation), and this shift leads to a large estimation error in precipitation. Fig. 6 shows the AMSR-E horizontal observation direction of the data corresponding to Fig. 5. From Fig. 6, it can be seen that the direction of the deviation coincides with the horizontal observation direction of AMSR-E.

Fig. 7 shows the distribution of the actual precipitation and the scattered precipitation of ice particles assuming the average height of ice particles of 14.0km. The assumption of 14.0 km was obtained from Table I, but even in the tropics, the 0 degree altitude is about 5 or 6 km, and convective precipitation exceeding 10 km is rare, so it is not very realistic.

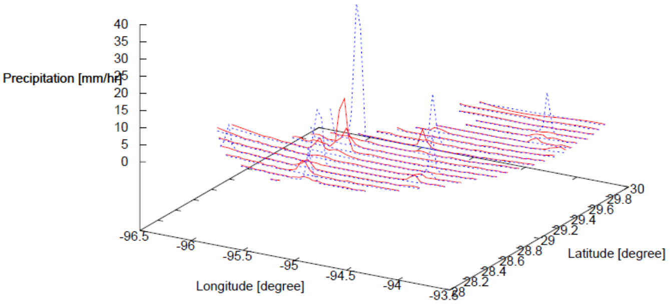


Fig. 5. Distribution of Actual Precipitation (Solid Line) and Estimated Precipitation (Doted Line) with Scattering Due to Ice Particles. Each Line Corresponds to Observation Scan.

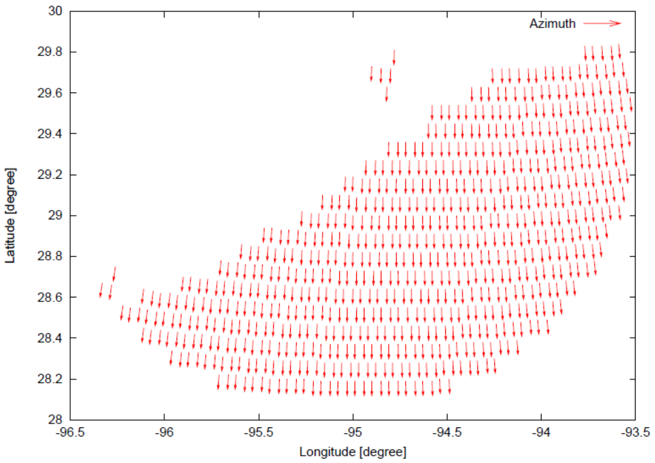


Fig. 6. Line of Sight Directions of AMSR-E.

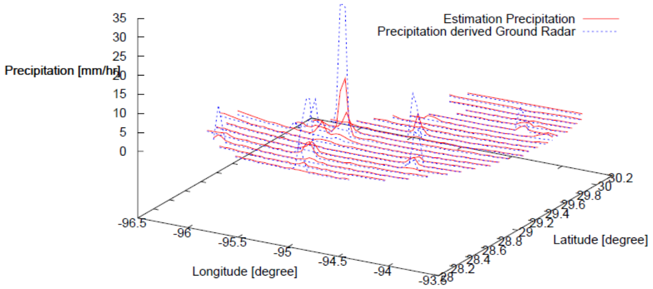


Fig. 7. Distribution of Actual Precipitation (Solid Line) and Estimated Precipitation (Doted Line) with Scattering Due to Ice Particles based on Assumption of mean Ice Particles Altitude. Each Line Corresponds to Observation Scan.

Therefore, since the estimation method based on this assumption includes all errors due to the falling speed of raindrops, etc.; it can be seen that valid evaluation is difficult. From Fig. 7, it can be seen that the distributions of the two precipitations are very similar, and the positions where the peaks of the precipitations appear coincide.

The validity of the assumption that the cause of the "shift" is the observation configuration was confirmed. Here, Fig. 5 shows the distributions of actual precipitation and ice particle scattering precipitation. The blue line is the distribution of actual precipitation [mm / hr], and the red line is the distribution of ice particle scattered precipitation [mm / hr]. Each line corresponds to a scan of the observation.

Fig. 6 shows the horizontal observation direction of the microwave radiometer AMSR-E. Furthermore, Fig. 7 shows the distribution of actual precipitation and scattered precipitation of ice particles assuming the average height of ice particles. The blue line is the distribution of actual precipitation [mm / hr], and the red line is the distribution of ice particle scattered precipitation [mm / hr]. Each line corresponds to a scan of the observation.

D. Consideration of Observation Configuration in Precipitation Estimation Method

If the average height of the ice particles can be assumed or known, the deviation can be corrected for the observation position of the 89 GHz data by using equation (2). In this study, as described above, the effect of the 18GHz data shift on the precipitation estimation accuracy is considered to be small, so it is not considered. Precipitation was estimated using the AMSR-E microwave radiometers at 18GHz and 89GHz when passing over the ground precipitation radars in Houston, July 2007 and Kwajalein, September 2007. At that time, the correction of the deviation by the observation configuration was added to the 89GHz data.

The specific procedure of consideration is shown below.

1) If the average ice particle height is unknown, estimate (assume) it by some method (this study uses the average ice particle height obtained in the previous section).

2) Calculate the deviation from the average height of ice particles, the incident angle of observation, the azimuth, etc.

3) Correct the deviation corresponding to the latitude and longitude information of the 89GHz data.

Also, when calculating equation (3), 18 GHz data and 89 GHz data at the same position are obtained, but their observation positions are shifted. Originally, all 89GHz data within the 18 GHz footprint should be used. However, when all the data in the footprint is used, even if the antenna pattern is taken into consideration, the effect of the correction of the position is hardly seen because the data is smoothed in the footprint. Therefore, if it is desired to correct the 89 GHz position effectively, it is better to use the nearest neighbor.

Therefore, in this study, the nearest 89 GHz data was used based on the position of the 18 GHz data. The precipitation estimated by these methods is called the estimated precipitation. Estimated precipitation and actual precipitation (surface precipitation measured by ground precipitation radar, etc.) were compared, and the correlation coefficient with Root Mean Square Error: RMSE was calculated.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_N (P_t - P_c)^2} \qquad (8)$$

RMSE of precipitation estimation for before correction (solid line) and after correction (Doted line) as a function of mean precipitation (Houston data) is shown in Fig. 8.

Also, RMSE of precipitation estimation for before correction (solid line) and after correction (Doted line) as a function of mean precipitation (Kwajalein data) is shown 9 in Fig. 9.

The results are shown in Table III (Houston) and Table IV (Kwajalein). In the table, the number of data indicates the number of AMSR-E constant areas for which precipitation was estimated, and Rt and Re are the precipitation (mm / hr) and the estimated precipitation (mm / hr), respectively, from the ground-based precipitation radar. Is shown. RMSE indicates the mean square error of both, and R indicates the correlation coefficient of both. The horizontal axis shows the actual rainfall in the observation area, and the vertical axis shows the RMSE of each data. Fig. 8 (Houston) and Fig. 9 (Kwajalein) show the relationship between average precipitation and RMSE.

Fig. 10 (Houston) and Fig. 11 (Kwajalein) show the average actual precipitation in the observation area on the horizontal axis and the correlation coefficient of each data on the vertical axis.

Fig. 8 shows the RMSE before correction (red) and after correction (green) of each observation data in Houston. The horizontal axis represents the average precipitation [mm / hr] of each data, and the vertical axis represents the RMSE [mm / hr] of the actual precipitation and the estimated precipitation. Fig.9 shows the RMSE of each observation data before correction (red) and after correction (green) at Kwajalein. The notation is the same as in the Houston diagram.

From Fig. 8, Fig. 9, Fig. 10 and Fig. 11, it can be seen that the RMSE is reduced and the correlation coefficient is increased by correcting the displacement. Therefore, it was confirmed that the accuracy of precipitation estimation was improved by the displacement correction. When the precipitation is relatively small, no improvement in accuracy is seen. Rain clouds with low precipitation are not sufficiently developed to produce ice particles, and thus the ice particles themselves may not be present. Moreover, it is difficult to estimate the average height of ice particles because the characteristics of the distribution of precipitation are not remarkable in a small amount of rain. For these reasons, the accuracy improvement when the precipitation is relatively small is considered to be insignificant.

Fig. 10 shows the correlation of each observation data in Houston before (red) and after (green). The horizontal axis is the average precipitation [mm / hr] of each data, and the vertical axis is the correlation coefficient between the actual precipitation and the estimated precipitation. Fig. 11 shows the correlation of each observation data at Kwajalein before (red) and after (green). The notation is the same as in the Houston diagram.

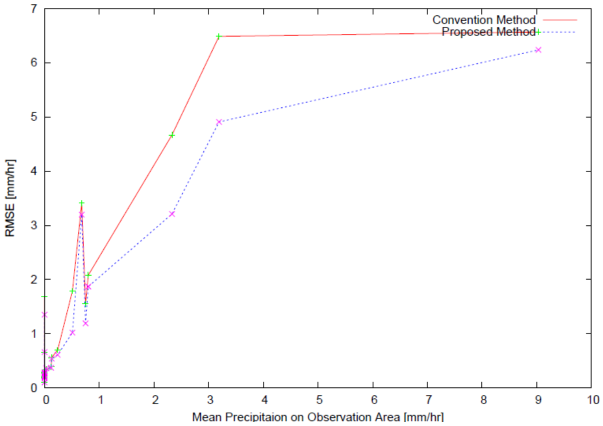


Fig. 8. Root Mean Square Error: RMSE of Precipitation Estimation for before Correction (Solid Line) and after Correction (Doted Line) as a Function of mean Precipitation (Houston Data).

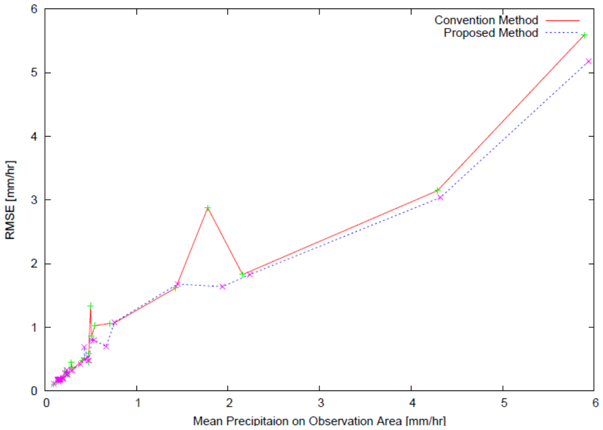


Fig. 9. Root Mean Square Error: RMSE of Precipitation Estimation for before Correction (Solid Line) and after Correction (Doted Line) as a Function of mean Precipitation (Kwajalein Data).

TABLE. III. RESULTS FROM PRECIPITATION ESTIMATION WITH/WITHOUT LOCATION DISPLACEMENT CORRECTION (HOUSTON DATA)

			Conventional Method				**Proposed Method**				
Data No.	**Date**	**No.of Data**	**Rt(mm/hr)**	**Re(mm/hr)**	**RMSE(mm/hr)**	**R**	**H(km)**	**Rt(mm/hr)**	**Re(mm/hr)**	**RMSE(mm/hr)**	**R**
1	7/2	676	0.24	0.4	0.7	0.42	14	0.25	0.38	0.61	0.61
2	7/4	627	9.03	7.85	6.57	0.69	10.5	9.01	7.19	6.24	0.75
3	7/5	599	0.51	0.3	1.79	0.38	3.5	0.36	0.29	1.02	0.47
4	7/6	741	0.8	0.87	2.08	0.5	13.5	0.79	0.82	1.87	0.64
5	7/7	1032	0	0.22	0.26	N/A	0	0	0.22	0.26	N/A
6	7/7	541	0	0.2	0.22	N/A	0	0	0.2	0.22	N/A
7	7/9	661	0	0.16	0.19	N/A	0	0	0.16	0.19	N/A
8	7/9	562	0	0.17	0.19	N/A	0	0	0.17	0.19	N/A
9	7/11	730	0	0.09	0.1	N/A	0	0	0.09	0.1	N/A
10	7/12	730	0	0.17	0.18	N/A	0	0	0.17	0.18	N/A
11	7/13	638	0	0.14	0.18	N/A	0	0	0.14	0.18	N/A
12	7/14	574	0	0.22	0.29	N/A	0	0	0.22	0.29	N/A
13	7/16	793	0.03	0.19	0.34	0.46	6	0.02	0.19	0.36	0.45
14	7/16	543	0	0.18	0.21	N/A	0	0	0.18	0.21	N/A
15	7/18	450	3.19	2.2	6.49	0.37	19	3.77	2.61	4.91	0.7
16	7/20	627	0.12	0.3	0.39	0.49	11	0.12	0.28	0.37	0.54
17	7/21	628	0	0.26	0.28	N/A	0	0	0.26	0.28	N/A
18	7/22	684	0	0.21	0.26	N/A	0	0	0.21	0.26	N/A
19	7/23	1091	0	0.41	0.65	0.36	12.5	0.01	0.44	0.66	0.35
20	7/23	543	0	0.14	0.18	N/A	0	0	0.14	0.18	N/A
21	7/25	676	0.13	0.5	0.56	0.64	8	0.13	0.49	0.54	0.66
22	7/25	471	0.75	0.55	1.55	0.65	11.5	0.6	0.46	1.19	0.65
23	7/27	684	2.33	2.94	4.67	0.6	17.5	2.44	3.08	3.21	0.84
24	7/27	712	0	1.06	1.69	0.16	0.5	0	0.9	1.35	0.18
25	7/28	720	0.68	0.53	3.41	0.18	2	0.66	0.44	3.2	0.21
26	7/30	568	0	0.16	0.19	N/A	0	0	0.16	0.19	N/A

TABLE. IV. RESULTS FROM PRECIPITATION ESTIMATION WITH/WITHOUT LOCATION DISPLACEMENT CORRECTION (KWAJALEIN DATA)

			Conventional Method				**Proposed Method**				
Data No.	**Date**	**No.of Data**	**Rt(mm/hr)**	**Re(mm/hr)**	**RMSE(mm/hr)**	**R**	**H(km)**	**Rt(mm/hr)**	**Re(mm/hr)**	**RMSE(mm/hr)**	**R**
1	9/1	931	0.75	0.56	1.06	0.61	11.6	0.77	0.56	0.98	0.68
2	9/1	939	0.29	0.12	0.38	0.66	8.6	0.29	0.12	0.32	0.76
3	9/3	909	0.51	0.36	0.86	0.51	9	0.51	0.36	0.8	0.58
4	9/3	1146	0.5	0.12	1.34	0.18	15.8	0.43	0.12	0.69	0.48
5	9/4	1070	0.14	0	0.17	0.12	11.2	0.14	0	0.17	0.12
6	9/5	902	0.22	0.11	0.32	0.37	12.2	0.21	0.12	0.27	0.56
7	9/5	174	0.29	0.1	0.45	0.49	13.4	0.24	0.1	0.35	0.7
8	9/6	404	0.2	0.01	0.21	0.2	2.6	0.2	0.01	0.21	0.21
9	9/6	1041	0.17	0	0.19	N/A	0	0.17	0	0.19	N/A
10	9/8	919	0.48	0.13	0.46	0.73	3	0.48	0.13	0.48	0.75
11	9/10	1002	0.55	0.42	1.03	0.7	9.8	0.54	0.42	0.8	0.82
12	9/12	897	0.24	0	0.27	0.3	17.2	0.24	0	0.26	0.29
13	9/12	956	0.16	0	0.17	0.16	2.8	0.16	0	0.17	0.1
14	9/13	1018	0.15	0.03	0.17	0.6	4.6	0.15	0.03	0.17	0.59
15	9/13	1231	0.17	0.02	0.18	0.22	8	0.17	0.02	0.19	0.22
16	9/15	1045	4.29	2.66	3.15	0.83	5.8	4.32	2.66	3.04	0.85
17	9/15	958	0.48	0.22	0.59	0.65	10.4	0.47	0.22	0.53	0.71
18	9/17	936	0.25	0	0.26	0.07	7.2	0.25	0	0.26	0.08
19	9/19	1000	5.89	2.84	5.59	0.7	15.2	5.94	2.84	5.18	0.79
20	9/20	1071	2.18	1.36	1.83	0.9	12.2	1.94	0.94	1.64	0.89
21	9/21	992	0.14	0.05	0.19	0.4	8.4	0.14	0.05	0.18	0.53
22	9/22	1003	1.43	0.72	1.62	0.71	9	1.45	0.72	1.68	0.72
23	9/22	1042	0.16	0.04	0.18	0.38	10.2	0.16	0.04	0.18	0.41
24	9/24	966	0.2	0.04	0.2	0.59	3.4	0.2	0.04	0.19	0.6
25	9/26	945	0.3	0.04	0.34	0.48	13	0.3	0.04	0.33	0.55
26	9/26	723	0.71	0.41	1.06	0.69	11	0.67	0.41	0.7	0.84
27	9/28	1130	0.43	0.12	0.52	0.63	10	0.43	0.12	0.49	0.69
28	9/28	936	0.41	0.13	0.48	0.62	9.2	0.38	0.13	0.42	0.73
29	9/29	381	0.15	0	0.15	N/A	0	0.15	0	0.15	N/A
30	9/29	915	0.1	0	0.11	N/A	0	0.1	0	0.11	N/A
31	9/30	204	1.78	0.52	2.88	0.33	20	2.24	1	1.83	0.8

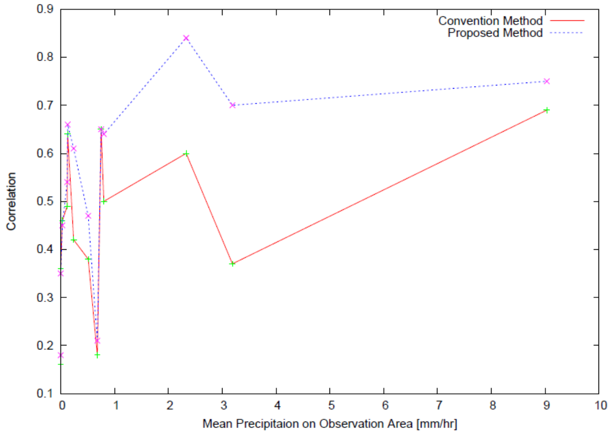


Fig. 10. Correlation Coefficient between Actual and Estimated Precipitation for before Correction (Solid Line) and after Correction (Doted Line) as a Function of mean Precipitation (Houston Data).

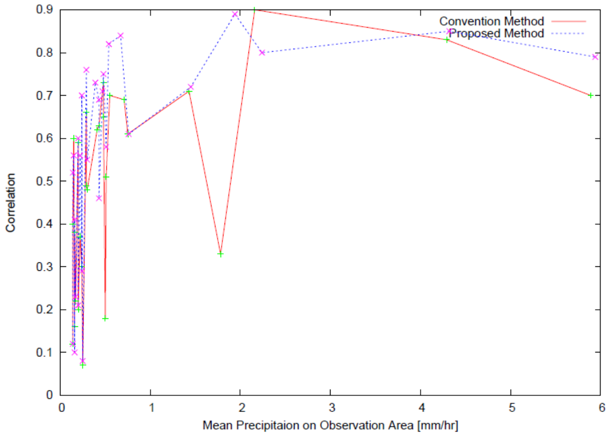


Fig. 11. Correlation Coefficient between Actual and Estimated Precipitation for before Correction (Solid Line) and after Correction (Doted Line) as a Function of mean Precipitation (Kwajalein Data).

Fig. 12 shows the Houston race on January 10, 2008. Indicates precipitation. Fig. 13 shows the precipitation estimated using AMSR-E data. On January 10th, heavy rainfall was observed locally, and convective clouds are thought to have occurred. Using the AMSR-E data on that day, the average ice crystal height was estimated to be 11 km. The RMSE was improved from 8.16 mm / hr to 6.57 mm / hr by the proposed method considering the ice crystal height with respect to the precipitation estimated by the conventional method, and the correlation coefficient was increased from 0.51 to 0.73.

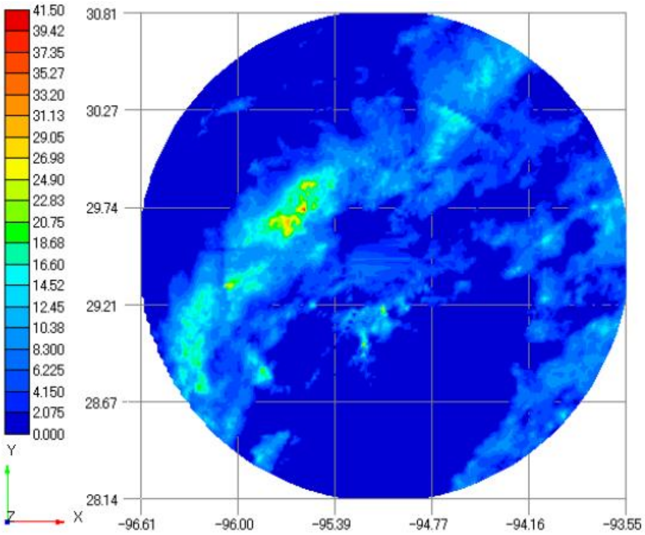


Fig. 12. Rainfall Rate Derived from Rain Radar which is Situated in Houston Measured on January 10 2008.

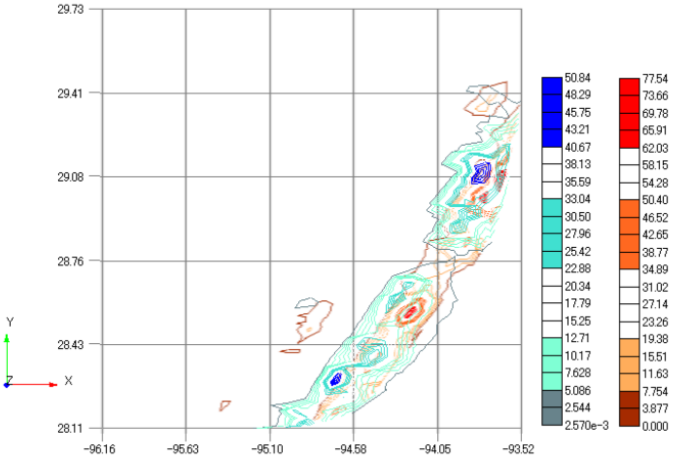


Fig. 13. Estimated Rainfall Rate with AMSR-E Data of Houston which is acquired on January 10 2008.

V. Conclusion

The method of considering the geometrical relationship between the observation target ice particles and the microwave radiometer in rainfall estimation was shown, and its validity was shown by comparing it with precipitation radar data on the ground. Observation objects that exist at high altitudes, such as ice particles, have a large difference between the observation point projected on the ground surface and the position of the upper troposphere where the observation object exists did. This effect was insignificant when the precipitation was small because ice particles were often absent, but the effect was significant when the precipitation was large.

In other words, the proposed method is effective and effective for AMSR data in Houston, which was shown as an example of a highly developed convective rain cloud with an ice cloud at the top. In the case of Kwajalein, the effect is insignificant. In addition, the proposed method requires an assumption of ice particle height, and it is necessary to make assumptions based on climatic values. In addition, microwaves in the 89 GHz band, which are considered to be sensitive to ice particles, are not only sensitive to ice particles, so it must be taken into account that they are also affected by the presence of non-ice particles.

The proposed method works well for rainfall rate estimation with the AMSR and AMSR-E data.

VI. Future Research Works

In this study, rainfall was estimated considering the geometric relationship between the observation target and the microwave radiometer. However, it takes a certain amount of time for an object at a high altitude to change (affect) the physical quantity on the ground surface. Therefore, when observing a fluctuating rain, it is necessary to consider this time difference. When observing very local and rapidly changing phenomena, such as cumulonimbus clouds, it is necessary to consider the spatial and temporal lags due to their observation configurations in order to perform more precise observations.

Acknowledgment

The author would like to thank Dr. Kenta Azuma of former student of Saga University and Prof. Dr. Hiroshi Okumura of Saga University for his valuable comments and suggestions.

References

[1] Roy W.Spencer and H.M ichael Goodman and Robbir E.Hood, Precipitation Retrieval over Land and Ocean with the SSM/I : Identification and Characteristics of the Scattering Signal,Journal of Atmospheric and ocean technology, vol.6, 254-273, April, 1989.

[2] Guosheng Liu and Judith A.Curry,Retrieval of Precipitation from Satellite Microwave Measurement Using Both Emission and Scattering, Journal of Geophysical reseach, vol.97, no.D9, 9959-9974, June 20, 1992.

[3] Guosheng Liu and Judith A.Curry and Rong- Shyang Sheu, Classification of cloud over the western equatorial Pacific Ocean using combined infrared and microwave satellite data,Journal of Geophysical reseach, vol.100, no.D7, 13, 811-13, 826, July 20, 1995.

[4] Guosheng Liu,Description Precipitation Retrieval Algorithm for ADEOS II AMSR, http://sharaku.eorc.jaxa.jp/AMSR/doc/alg/7alg.pdf,2002.

[5] Roberti, L., Haferman J., and Kummerow C. Microwave radiative transfer through horizontally inhomogeneous precipitating clouds. Journal Of Geophysical Research. 99, 16, 707-16, 718. (1994).

[6] Kummerow, C., P.Poyner, W.Berg and J. Thomas-Stahle, The E□ects of Rainfall Inhomogeneity on Climate Variability of Rainfall stimated from Passive Microwave Sensors, JAOTEC, 21, 624-638, 2004.

[7] Takao Takeda, Science of Water Cycle-Behavior of Clouds-Tokyodo Shuppan、1987.

[8] Kohei Arai, T. Igarashi and C. Ishida, Evaluation of MOS-1 Microwave Scanning Radiometer(MSR) data in field experiments, Proc. of the 18th International Symposium on Remote Sensing of Environment, 1-8, 1984.

[9] Y. Itoh, K. Tachi, Y. Sato and Kohei Arai, Advanced Microwave Scanning Radiometer: AMSR, Preliminary study, Proc. of the IGARSS'89, I1-4, 273-276, 1989.

[10] K. Tachi, Kohei Arai and Y. Satoh, Advanced Microwave Scanning Radiometer -Requirements and Preliminary Design Study-, IEEE Trans. on Geoscience and Remote Sensing, Vol.27, No.2, pp.177-183, Jan.1989.

[11] Kohei Arai and K. Teramoto, Antenna Pattern Correction and SST Estimation Algorithms for AMSR Proceedings of the AMSR Science Workshop, (1997).

[12] Kohei Arai, Advanced Microwave Scanning Radiometer(AMSR) Proceedings of the AMSR Science Workshop Tokyo, Japan, 1997.

[13] Kohei Arai and Jun Sakakibara, Simultaneous estimation of sea surface temperature, wind speed and water vapor with AMSR-E data based on improved simulated annealing, Proceedings of the Renewable Energy Resources Symposium, 00547, 2006.

[14] M. Konda, A. Shibata, N. Ebuchi and Kohei Arai, Correction of the effect of relative wind direction on wind speed derived by AMSR, Journal of Oceanography, 64, 395-404, 2006.

[15] Kohei Arai, Kenta Azuma, Precipitation estimation using AMSR data considering geometric relationship between observation target and radiometer, 49,1,32-40,2010.

[16] Kohei Arai, Nonlinear Optimization Based Sea Surface Temperature: SST Estimation Methods with Remote Sensing Satellite Based Microwave Scanning Radiometer: MSR Data, International Journal of Research and Reviews in Computer Science (IJRRCS) Vol. 3, No. 6, 1881-1886, December 2012, ISSN: 2079-2557, 2012.

[17] Kohei Arai, X.Liang, Q.Liu, Method for estimation of rain rate with Rayleigh and Mie scattering assumptions on the Z-R relationship for different rainfall types, Advances in Space Research, 36, 5, 813-817, 2005.

[18] Kohei Arai, Comparison between Rayleigh and Mie scattering assumptions for Z-R relation and rainfall rate estimation with TRMM/PR data, International Journal of Advanced Research in Artificial Intelligence, 2, 8, 1-6, 2013.

[19] Arai,K., Detecting Algorithm for Rainfall Area Movement based on Kalman Filtering, Proceedings of the NSAT/SWT Symposium, Kyoto, Nov. 1995.

[20] Kohei Arai and Masanori Saka, Reconstruction of cross section of rainfall situations with precipitation radar data based on wavelet analysis, Abstracts of the 35th Congress of the Committee on Space Research of the ICSU, A1.1-0232-04, (2004).

AUTHOR'S PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 37 books and published 570 journal papers. He received 30 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.html.

Image-based Individual Cow Recognition using Body Patterns

Rotimi-Williams Bello[1], Abdullah Zawawi Talib[2], Ahmad Sufril Azlan Mohamed[3]
Daniel A. Olubummo[4], Firstman Noah Otobo[5]
School of Computer Sciences, Universiti Sains Malaysia, 11800, Pulau, Pinang, Malaysia[1, 2, 3]
Department of Computer & Information Systems, Robert Morris University, Moon-Township, Pennsylvania, USA[4]
Department of Mathematical Sciences, University of Africa, Toru-Orua, Bayelsa State, Nigeria[1, 5]

***Abstract*—The existence of illumination variation, non-rigid object, occlusion, non-linear motion, and real-time implementation requirement has made tracking in computer vision a challenging task. In order to recognize individual cow and to mitigate all the challenging tasks, an image processing system is proposed using the body pattern images of the cow. This system accepts an input image, performs processing operation on the image, and output results in form of classification under certain categories. Technically, convolutional neural network is modeled for the training and testing of each pattern image of 1000 acquired images of 10 species of cow which will pass it through a series of convolution layers with filters, pooling, fully connected layers and softmax function for the pattern images classification with probabilistic values between 0 and 1. The performance evaluation of the proposed system for both training and testing data was carried out for each cow's identification and 92.59% and 89.95% accuracies were achieved respectively.**

Keywords—Cow; body patterns; convolutional neural network; image; recognition

I. Introduction

Cows in the past were classically monitored with the sole aim of aiding tracking, health information, performance recording, prevention against manipulation and swapping, and verification of false insurance claims. There are basically two recognition techniques employed for the identification of the animal. One recognition technique leaves a permanent mark on the animal for identification while the other recognition technique leaves a temporary mark. Examples of the recognition technique that leaves a permanent mark are found in [1], [2], [3], [4], [5] with their drawbacks. The tattooing of ears, tagging of ears, microchips implant and branding are popular invasive identification techniques that leave a permanent mark on the animal's body with so many challenges such as animal infections, mild sepsis, and hemorrhaging [2], [3].

Examples of the recognition technique that leaves a temporary mark on the body of the animal for identification purposes are found in the work of Barron et al. [6] with their drawbacks. Among the classical methods of animal identification are drawing, tagging, tattooing, branding, notching, and Radio Frequency Identification (RFID). However, classical methods of animal identification have notable adoption problems which have contributed to the low acceptance rate of the methods among the cow breeders. The classical methods of animal identification are not reliable; they are prone to fraudulent activities such as swapping, duplication and forgery of the so called unique identification numbers tagged on the animal's body [7], [8], and therefore cannot meet the required level expected from them for the monitoring and identification of animal [9].

Many automatic systems have been proposed recently for the monitoring and identification of cow however, most of these devices are sensor based and sometimes become burden and injurious when worn on the body of the animal [10]. There is need for automatic cow monitoring system in livestock farm to be developed as there is uprising in the number of cow year in year out in almost every part of the world and there is great task involved in monitoring cow manually. Lu et al. [11] proposed cow traceability system that was based on the iris analysis for the enhancement of cow management. The image quality assessment of the captured iris sequences was firstly made before the clear iris was selected. By using segmentation that was based on edge detection, the inner and outer boundaries of the iris of the cow were fitted as ellipse form. The iris image was normalized using geometric method and both the local and the global features of the iris of the cow were extracted using 2D complex wavelet transform. However, in an unconstrained environment where there is greater possibility of getting poor quality image of cow's iris, this method may not be appropriate for a reliable traceability.

By using video data, there is every possibility that the problems attributed to the classical methods can be mitigated using the visual based automatic cow recognizing system. The recognition of individual cow in the automatic cow monitoring implementation process enables behavior monitoring of individual cow at long run for body condition score which plays important role in the health condition of individual cow. The system proposed in this paper is image-based individual cow recognition using body patterns. The rest of the paper is organized as follow. Presented in Section 2 are the literature review, followed by the material and method in Section 3, the results and discussion are in Section 4. The conclusion is in Section 5.

II. Literature Review

The conventional constructs of identifying animal can be categorized into: (1) permanent recognition construct (PRC);

(2) semi-permanent recognition construct (SRC); and (3) temporary recognition construct (TRC); [12], [13]. The tattooing of ear and body, tagging of ear, microchip implants and branding are referred to as PRC recognition methods [14] but with several limitations [15] such as: (1) lack of large scale production of various metal clips and plastic tags that can be enough for the identification of large-scale animal; (2) easy lost of the available ear tags due to ear tearing; (3) infections of animals such as cattle and other ruminant animals due to notches [16], [17], [18], moreover, more than half percentage of the animals are infected from the injury sustained on their ear due to the implanted plastic ear-tags, reason being that, the ear-tags cause various health challenges such as local inflammation, thickening of the flesh, presence of pus-forming bacteria, and loss of blood through the notch [17], [13]. Cattle recognition using methods such as pattern sketching and collar is SRC method. Furthermore, the use of dye or paint and radio frequency identification (RFID) based recognition are referred to as TRC for the recognition of animal [12], [19].

According to [20], the sketching pattern is applied for the recognition of cattle such as Holsteins and Guernsey with broken color. High drawing skills of an individual for sketching is needed which should be comparable to standard image quality and positively affect the cattle identification process. However, this method cannot be used for the identification of solid collared breeds such as Red Poll and Brown Swiss breed as some artificial marking methods such as ear tagging and tattooing that are discrimination based are needed. However, the method of ear tagging damages the cattle's ear at the long run. As iterated in Petersen's work [21], muzzle print-based cattle recognition method using blue ink and A-5 paper [22] was the first attempt to get permanent recognition method for cattle. In the method, skills are required to acquire the muzzle pattern's print image, by holding firm the cattle.

Lately, the research community has shifted attention to advancing cattle recognition using image of muzzle print as a new paradigm for cattle identification [22], [20]. According to [23], print image of muzzle pattern is made up of beads and ridges patterns. Muzzle dermatoglyphics such as granola, ridges, and vibrissae from various breeds are not the same [16]. Similarly, proposed in Mishra et al. [24] is method of cattle breeds recognition using the beads and ridges features of muzzle print images. Similar to the work of Mishra et al. [14] is Minagawa et al. [22], they proposed a cattle identification method using muzzle print, the performance evaluation was made using filtering techniques for muzzle image analysis and morphological approaches. Equal Error Rate (EER) of 0.419 was reported by them.

Contrary to Minagawa et al. [22] is a framework proposed by Barry et al. [25]. The framework is a cattle recognition using muzzle print images. They reported the 241 false non-match rates (FNMR) over 560 genuine acceptance rate (GAR) and 5197 false matches over 12,160 impostors matching closely with the same value of EER of 0.429, respectively. In their cattle identification effort, Kim et al. [26] proposed a method that could recognize the Japanese black cattle using the cattle face's pixel intensity [26]. Proposed in [27] is a local binary pattern (LBP) based model for recognition of cattle using the texture features of cattle facial representation. Proposed in [28] is an approach for cattle recognition based on Speeded Up Robust Feature (SURF) descriptor. The approach was an enhancement of Petersen's method for cattle identification. The results of experiment was reported based on the image datasets of 4 cattle breeds used which were captured on A-5 paper with blue inked for the purpose of cattle recognition. Proposed in [20] is a matching refinement technique in scale invariant feature transform (SIFT) descriptor for cattle recognition using database of 160 muzzle print images. By the application of matching refinement technique in SIFT approach, the matching scores of the key-points of muzzle print images were computed. Nevertheless, the performance of the matching refinement approach and the original SIFT approach were compared, and the value of EER equal to 0.0167 was achieved.

Proposed in Awad et al. [29] is a framework for recognizing cattle using SIFT descriptor approach. The approach is used for localizing and detecting the beads and ridges' key points in the images of muzzle print for the cattle identification. The RANdom SAmple Consensus (RANSAC) technique incorporated in the SIFT algorithm is used for the palliation of the outliers in muzzle image for an improved, robust, and reliable cattle identification. Database of 90 muzzle images was used for the experiment where 15 muzzle images were captured from each cattle of 6 in number. Proposed in Tharwat et al. [23] is an approach of cattle recognition that was based on muzzle image using the technique of local texture descriptor. The technique works in such a way that the texture extraction algorithm that was based on local binary pattern used the local texture features extraction from the images of muzzle point. The involvement of more processing time in the cattle recognition process is a major limitation of the technique.

Object recognition method that is based on CNN was proposed in [30]. The proposed architecture which combines RGB image and its corresponding depth image for object recognition is made up of two unconnected CNN processing streams, which are sequentially integrated with a late fusion network. ImageNet [31] is employed for the training of the CNNs in which the depth image is encoded as a rendered RGB image, making the information that is contained in the depth data to go round over all the three RGB channels, and subsequently, a standard and pre-trained CNN is employed for the recognition. Due to limited availability of large scale depth datasets that are labeled, CNNs that are pre-trained on ImageNet [32] are employed. Proposed in [33], is another object recognition method, which employs deep CNN. The proposed method also uses CNN, which is pre-trained for image classification and provides a robust, semantically meaningful feature set. The depth information is integrated by rendering objects from a canonical perspective and getting the depth channel colorized according to distance from the object center.

Jingqui et al. [34] proposed the method of object recognition based on image entropy; this was aimed at identifying the behavior of cow object that is on the motion against a complicated background. They used the minimum bounding box and contour mapping for the real-time capture of behavioral and characteristic features displayed by the cow. Although the approach used has time-saving advantage for cow breeders and yields a high recognition rate of estrous and hoof-disease not less than 80%, the time correlation of cow behaviors was not integrated.

Andrew et al. [35] demonstrated the suitability of computer vision pipelines that utilize deep neural architectures to carry out automated Holstein Friesian cattle detection in addition to individual identification in a farm set up. They showed that it is possible to perform robustly Friesian cattle detection and localization with an accuracy of 99.3% on the available dataset. Although they showed the capability of their method in the scenarios presented, they did not consider complicated setups such as faster moving, larger herds and tight animal gatherings.

In the process of extracting features from an image, Kumar et al. [36] posited that pre-processing is important for object tracking accuracy but feature extraction and representation algorithms that are based on appearance are unable to perform the recognition of object as a result of image blurriness due to noise, low illumination and the unconstrained environment under which the images were captured. Therefore a method based on feature descriptor techniques is utilized for the unique identification of individual object. Based on the pre-processing process, reliable results were obtained from the tracking process of the object. Pre-processing which majorly involves particle filtering and segmentation of muzzle point images is necessary in the features extraction process. The primary aim of undergoing pre-processing of the muzzle images using enhancement algorithms before the feature extraction and matching process take place is to ensure that the muzzle images are enhanced before the analysis of the extracted texture features and for better representation in the feature space.

III. Material and Method

A. Equipment for Experiment

Ten (10) species of cow were examined in recognizing the characteristic of individual cow, each having 100 images making 1000 images in total. The patterns of the black and white body of the 10 species of cow were used for the calculation of the input parameters values for training. 400 images of body patterns (40 cows (subject) × 10 images of each subject) were used for the training of the proposed deep learning approach in the training phase. 600 pairs of testing (60 cows (subject) × 10 images of each subject) of the body patterns images in each fold were used for testing the probe images in the testing phase. By middle of September 2018, a test was performed in order to get the image data and the image data was analyzed accordingly by image processing. A charged coupled device (CCD) camera was employed for the side image capture of each cow. In order to obtain images of required width (235-270cm), the CCD camera was placed on a high pole away from the experimental system centerline. The image processing system was strategically placed in a location through which the cow passed everyday with minimized illumination variation for the production of noiseless and clear images as shown in Fig. 1.

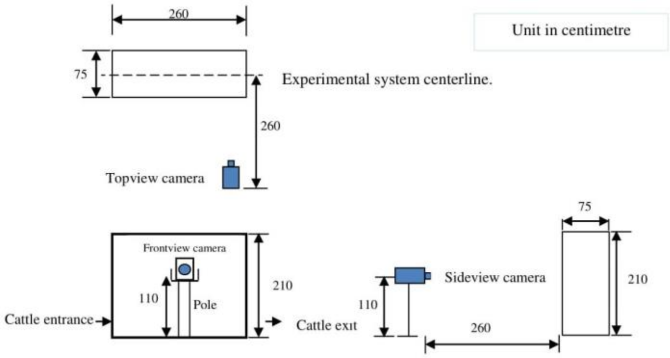


Fig. 1. Dimensional Sketch of the Individual Cattle Recognizing System.

The cow recognition and identification system can run on any Windows-based personal computer. A faster computer system is recommended for the processing of the images that involves calculations and processing on the go. The personal computer specifications for development of the cow recognition and identification system are Intel core i5 Processor, 8 Gigabyte of RAM, Graphics card, 2 terabytes of hard disk space, a CCD digital camera, and a computer monitor for digitizing, displaying, and processing multiple images. The specification for the execution of the image-processing and computer vision elements is OpenCV and its library.

B. Processing of Images

The filtration technique used for this work is Gaussian filtering technique while multi-layer deep learning neural network was used as a classifier for the cow identification and contrast limited adaptive histogram equalization (CLAHE) was used for enhancement of the contrast between the cow's body patterns. The difference of the Gaussian filter was got by finding the difference between two Gaussian functions [37].

Fig. 2 shows some image samples of cow's body patterns from the database. Fig. 3 shows the database containing blurred image patterns of the cow's body affected by the unconstrained environment and postures of the cow leading to poor quality of the images. Using Norouzzadeh et al. [38], we filtered the images to get rid of the blurriness, background patches and low illumination.

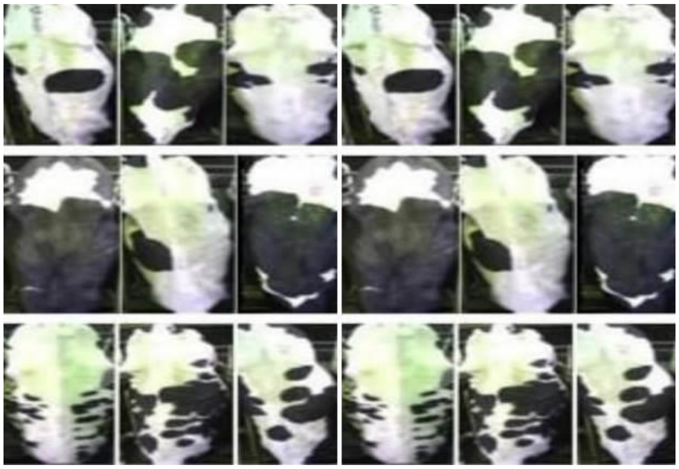


Fig. 2. Images of Cow's Body Patterns from the Database.

Fig. 3. Blurred and Poor Illumination Images of Cow's Body Patterns.

In order to enhance the identification process and remove the patches and the noises from the captured images that were collected, various image pre-processing techniques were applied. Low illumination and poor image quality are the most two fundamental challenges confronting image acquisition especially images of cow's body patterns. The images captured in an unconstrained environment were converted to grayscale images in order to reduce the patches and the noises captured with them. The converted images were improved upon by contrast limited adaptive histogram equalization based image processing technique.

The pre-processing technique accepts the images in their color form and converts them to grayscale before fetching them into the filter for removal of the patches and the noises contained in the captured images. The feature extraction involves the convolution and pooling operations on the images until the images get to the classifier for classification analysis for the generation of the desired output (Fig. 4). The removal of the noises was carried out using an auto-encoding technique. Stacked denoising auto-encoder (SDAE) technique initializes deep network and it is applicable for encoding and decoding the texture features of the image patterns that were extracted and encoding the extracted sets of features for optimum representation of the feature [39].

Technically, convolutional neural network (CNN) is modeled for the training and testing of each input image which will pass it through a series of convolution layers with filters, pooling, fully connected layers and softmax function for the image classification with probabilistic values between 0 and 1. As shown in Fig. 4, the first layer to extract features from the input image is convolution. Convolution primarily conserves the relationship between pixels by learning the image features using squares of input data. It involves a mathematical operation with two inputs such as image matrix and a filter. When there are too large images, pooling layers primarily reduce the number of parameters (dimensionality size). In the proposed CNN as seen in Fig. 4, the operation of the pooling is applied individually to each feature map.

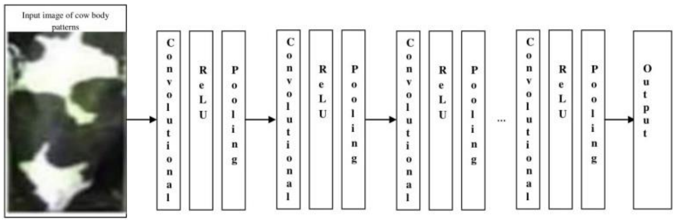


Fig. 4. Neural Network with Convolutional Layers for Cow Recognition.

Generally, the more the convolutional steps become, the more the complex features possibility of being recognized becomes using the proposed network. Until the system can dependably recognize objects, the whole process is repeated in successive layers. Each layer's neurons of the CNN as seen in Fig. 4 are in 3D arrangement, making a transformation of a 3D input to a 3D output. For instance, for an input image, the first layer which is the input layer takes the images as 3D inputs, with height, width and color channels as the dimensions of the image. The first convolutional layer's neurons connect to the input images' regions and change them into a 3D output. Each layer hidden units learn nonlinear combinations of the original inputs which becomes the inputs for the layer that follows. By this, at the end of the network, the learned features become the inputs to the classifier.

The intensity values of the gray scale of the background images are more than 100 but less than 150 in respect to the colors of the cows' body surface. 128 was fixed as the pixel's threshold value for the whole image. While 1 is assigned as the binary values for the intensities that are greater than the threshold value of 128, 0 is assigned as the binary values for the intensities that are less than the threshold value of 128. Because the threshold value could be changed with illumination and noise, it becomes very important. Individual cow's image is captured for the identification of their individual characteristics. Individual cow identification using unique body patterns is made possible because of the invariant of the body patterns to growth. This uniqueness enables the patterns to be used as the input layer values in the neural network algorithm.

IV. Results and Discussion

Having tried out the effectuality of the proposed approach using images of cow's body patterns for the recognition and identification of the cow, the comparison with other recognition algorithms is attained in order to evaluate the accuracy of the identification in proliferation settings. Evaluating the performance of the experimental results, the database of the cow's body images is segmented as follows: (1) the training phase; and (2) the testing phase. 400 body images of different subjects (40 cows (subject) × 10 images of each subject) were used for the training of the proposed approach in the training phase. 600 pairs of testing (60 cows (subject) × 10 images of each subject) of the body patterns images in each fold were used for testing the probe images in the testing phase.

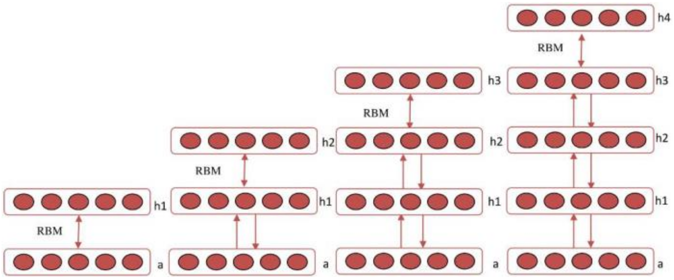


Fig. 5. RBM-based DBN Model.

For the training of the proposed deep learning framework using deep belief network (DBN) as shown in Fig. 5, there is a need for a monolithic database amount. Although the number of cow's body images in the database is encouraging, it is not satisfactory enough to train the stacked denoising auto-encoder with a database of 1000 worth of cow's body patterns images. Therefore a transfer learning approach is needed to fine-tune the weight between the input and the hidden layer and determine the pre-training of the proposed deep learning approach.

The basic mathematical steps that are involved in using the deep belief network for this work are as follows:

Problems setting: Given a training set of pre-processed body pattern image data

$$T = \{(x_1y_1),(x_2y_2),\dots,(x_Ny_N)\} \quad (1)$$

of which $(x_1y_1), i = 1,2,\dots,N$ denotes the sample point, $x_i \in X \subseteq R^n$ is the sample image data while $y_i \in Y$ is the corresponding tag of the label; the recognition procedure of proposed system is to input data set T to the network, find the mapping between input X and output Y to form a generative joint probability distribution model formula $P(X,Y)$, generate the output y_{N+1} by

$$y_{N+1} = \underset{y_{N+1}}{\arg max}\, \hat{P}(y_{N+1}|\, x_{N+1}) \quad (2)$$

for a given prediction sample x_{N+1}, and judge the image classification of x_{N+1} according to y_{N+1}. The system contains the following parts as shown in Fig. 4:

The proposed cow's body patterns image identification using deep belief network and a back propagation (BP) network layer, wherein the multi-layer RBM is used to input data feature learning to achieve abstraction and dimensionality reduction of data through the hierarchical feature learning is as shown in Fig. 5; BP network layer is a categorical network, and it is to categorize the abstracted higher-level features through softmax function. The softmax function, also known as softargmax or normalized exponential function, is a function that takes as input a vector of K real numbers, and normalizes it into a probability distribution consisting of K probabilities proportional to the exponentials of the input numbers.

The first part of the processes as shown in Fig. 4 is "preprocessed cow's body patterns images" which are introduced as inputs to the proposed networks for features extraction and classification.

The second part is "pre-training." For a given training set of image data $= \{x_1, x_2, \dots, x_N\}$, the learning system obtains a model through learning (or training) to describe the mapping relationship between input and output variables. This work assumed that RBM model has this descriptive ability, therefore it consists of several layers, through which the input is the image expression data vector while the output is the abstracted higher-level feature vector. Each layer of RBM networks undergoes individually unsupervised training to ensure that feature information is preserved to the uttermost as feature vectors are mapped to different feature spaces. To construct the joint distribution model of visible layer and the hidden layer through energy function, the joint probability maximum likelihood of training sample under model parameter $\hat{\theta}$ is calculated by

$$P(v|\hat{\theta}) = \frac{1}{Z(\theta)}\sum_h e^{-E(v,h|\theta)} \quad (3)$$

The third part is "fine tuning." Fine-tuning is a common strategy in deep learning to carry out supervised learning through tagged sample training set $T = \{(x_1^{'}, y_1), (x_2^{'}, y_2), \dots, (x_N^{'}, y_N)\}$. After that, the top feature vectors corresponding to sample output by the multi RBM network are formed based on the training set of statistical classification structure. This part is a BP network; it takes a specific dimension feature vector to a softmax function. In order to get the best connection weights, this work considered solving the following optimization problem using particle swarm optimization (PSO), so that the loss of function in the training set is minimized.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \frac{1}{N} l(\theta, x_n, t_n) \quad (4)$$

The last part is the "class identification." Tested sample x_{N+1} as network input is subjected to feature learning and abstraction through a network model training to produce a corresponding output y_{N+1} by

$$y_{N+1} = argmax\hat{P}(y_{N+1}|x_{N+1}) \quad (5)$$

and thus achieve classification.

For the evaluation of performance, the local feature descriptor technique was used to extract and encode texture features of the cow's body patterns. As earlier mentioned, the normalization and the descriptor process help in mitigating the external factors such as low illumination, poor image quality, and background patches affecting the captured images. In performing the tasks involved in this process, cells are converted to blocks. During this process, blocks are overlapped and cells shared among the blocks and normalized separately. Scale-invariant Feature Transform (SIFT) and Rectangular-Histogram of Oriented Gradients (R-HOG) are similar though, they don't align to their dominant orientation (Fig. 8(b)). SDAE produced the best experimental results (Fig. 6 and Fig. 7) when compared to other approaches used in this work making it fit the most for the denoising. 400 body images equivalent to (40 cows (subject) × 10 images of each subject) were chosen randomly for system training and 600 body images equivalent to (60 cows (subject) × 10 images of each subject) were used for the testing. The experimental results are reported and analyzed as found in Table I.

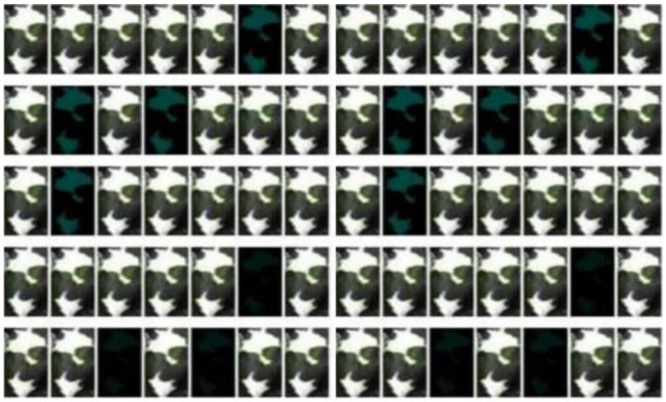


Fig. 6. Illustration of 17% Corrupted Images of Cow's Body Patterns using SDAE.

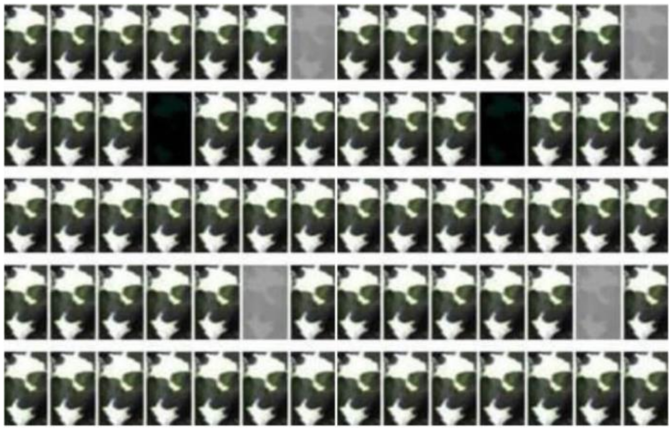
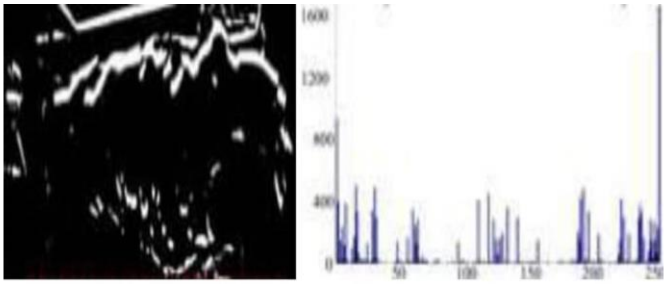


Fig. 7. Illustration of 5.7% Corrupted Images of Cow's Body Patterns.



(a) (b)

Fig. 8. (a) Binary Pattern of Cow's Body; (b) Histogram.

As it is shown in Table I, the evaluation of the system performance was carried out on the cropping, the training data, and the testing data for the overall achievement of the research objective. The average cropping accuracy of the captured video data is 79.45%, and the identification accuracy of the training data is 92.59% with the testing data having the identification accuracy of 89.95%.

TABLE. I. RECOGNITION ACCURACY (%)

Cropping	79.45%
Training data	92.59%
Testing data	89.95%

The significant reason for binary patterns (Fig. 8(a)) is to sum up the local structure in a block through comparison of each pixel with its neighborhood [40]. Each pixel coded with a sequence of bits is colligated with the connection between the pixel and one of its neighbors. The center pixel's intensity is denoted with 1 if it is greater than or equal to its neighbor, and denoted with 0 if otherwise with a binary number at the end created for each pixel.

V. CONCLUSION

Image-based individual cow recognition using body patterns was the main work carried out in this research. Cows usually are identified to prevent them from being stolen or protect them from danger, and in many agricultural settings, their behaviors are usually studied using imaging technology to enable timely monitoring and identification of health challenges. CNN and some other popular image recognizing techniques such as DBN, SDAE, CLAHE, Gaussian filter, binary pattern, were employed in this work for the cow recognition. The various techniques were discussed in details as they are applicable to the cow recognition process. Datasets of 1000 images of cow's body patterns from 10 species of cow were created for this work where 400 images were employed for the training and 600 images were used for the testing. The advantage of using this datasets is the various species of cow whose images are contained in the database used for the recognition. Gaussian filtering technique was used as the filtration technique; this was supported by SDAE for denoising while multi-layer convolutional neural network was used as a classifier in comparison to deep belief network which needs a monolithic database amount for the cow identification, and contrast limited adaptive histogram equalization (CLAHE) was used for enhancement of the contrast between the cow's body patterns. The performance evaluation of the proposed system for both training and testing data was carried out for each cow's identification and 92.59% and 89.95% accuracies were achieved respectively. Although this work has been able to apply modern image-based identification method for the recognition of cow using body patterns, recognition of occluded and non-linear moving object such as cow in real-time using the object's multi-features is a work that we consider worthy of investigating in the future.

ACKNOWLEDGMENT

This work was supported in part by the Institute of Postgraduate Studies, Universiti Sains Malaysia, 11800 USM, Penang, MALAYSIA.

REFERENCES

[1] Z. Miao, K. M. Gaynor, J. Wang, Z. Liu, O. Muellerklein, M. S. Norouzzadeh., and W. M. Getz, "Insights and approaches using deep learning to classify wildlife", Scientific reports, vol. 9, no. 1, pp. 1-9, 2019.

[2] S. Kumar, and S. K. Singh, "Cattle Recognition: A New Frontier in Visual Animal Biometrics Research", Proceedings of the National Academy of Sciences, India Section A: Physical Sciences 1-20, 2019.

[3] T. T. Zin, C. N. Phyo, P. Tin, H. Hama, and I. Kobayashi, "Image technology based cow identification system using deep learning", Proceedings of the International MultiConference of Engineers and Computer Scientists 1, 2018.

[4] R. W. Bello, and O. M Moradeyo, “Monitoring Cattle Grazing Behavior and Intrusion Using Global Positioning System and Virtual Fencing,”, Asian Journal of Mathematical Sciences, vol. 3, issue 4, pp. 4-14, 2019.

[5] R. W. Bello, “An overview of animal behavioral adaptive frightening system,” International Journal of Mathematics and Physical Sciences Research, vol. 6, issue 1, pp. 126-133, 2018.

[6] U. G. Barron, F. Butler, K. McDonnell, and S. Ward, “The end of the identity crisis? Advances in biometric markers for animal identification”, Irish Veterinary J., vol. 62, no. 3, pp. 204–208, 2009.

[7] M. Shen, L. Liu, L. Yan, M. Lu, W. Yao, and X. Yang, “Review of monitoring technology for animal individual in animal husbandry”, Nongye Jixie Xuebao = Transactions of the Chinese Society for Agricultural Machinery, vol. 45, no. 10, pp. 245-251, 2014.

[8] R. W. Bello, and S. Abubakar, “Development of a Software Package for Cattle Identification in Nigeria,” Journal of Applied Sciences and Environmental Management, vol. 23, no. 10, pp. 1825-1828, 2019. DOI: https://dx.doi.org/10.4314/jasem.v23i10.9.

[9] R. W. Bello, A. Z. H. Talib, and A. S. A. B. Mohamed, “A Framework for Real-time Cattle Monitoring using Multimedia Networks”, International Journal of Recent Technology and Engineering, vol. 8, issue 5, 2020.

[10] D. Grooms, “Radio Frequency Identification (RFID) Technology for Cattle”, Extension Bulletin E-2970, Michigan State University, Jan. 2007.

[11] Y. Lu, X. He, Y. Wen, and P. S. Wang, “A new cow identification system based on iris analysis and recognition”, In International Journal of Biometrics, vol.6, no.1, pp.18-32, 2014.

[12] S. Kumar, S. Tiwari, and S. K. Singh, “Face recognition of cattle: can it be done?”, Proceedings of the National Academy of Sciences, India Section A: Physical Sciences, vol. 86, no. 2, pp. 137–148, 2016.

[13] S. Kumar, S. Tiwari, and S. K. Singh, “Face recognition for cattle. Third IEEE International Conference on Image Information Processing (ICIIP), pp. 65–72, 2015.

[14] J. Marchant, “Secure animal identification and source verification. JM Communications, UK, pp. 1-28, 2002.

[15] A. Allen, B. Golden, M. Taylor, D. Patterson, D. Henriksen, and R. Skuce, “Evaluation of retinal imaging technology for the biometric identification of bovine animals in northern Ireland”, Livest Sci., vol. 116, no. 1, pp. 42–52, 2008.

[16] A. Baranov, R. Graml, F. Pirchner, and D. Schmid, “Breed differences and intra-breed genetic variability of dermatoglyphic pattern of cattle”, J Anim Breed Genet., vol. 110, no. 1–6, pp. 385–392, 1993.

[17] A. Johnston, and D. Edwards, “Welfare implications of identification of cattle by ear tags”, The Veterinary Record, vol. 138, no. 25, pp. 612–614,1996.

[18] D. D. Wardrope, “Problems with the use of ear tags in cattle”, The Veterinary record, vol. 137, no. 26, pp. 675-675, 1995.

[19] Z. Wang, Z. Fu, W. Chen, and J. Hu, “A RFID-based traceability system for cattle breeding in china”, Proceedings of 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), vol. 2, pp. V2–567, 2010.

[20] A. Noviyanto, and A. M. Arymurthy, “Beef cattle identification based on muzzle pattern using a matching refinement technique in the sift method”, Comput Electron Agric., vo. 99, pp. 77–84, 2013.

[21] W. E. Petersen, “The identification of the bovine by means of nose-prints”, Journal of dairy science, vol. 5, no. 3, pp. 249-258, 1922.

[22] H. Minagawa, T. Fujimura, M. Ichiyanagi, K. Tanaka, and M. Fangquan, “Identification of beef cattle by analyzing images of their muzzle patterns lifted on paper”, Proceedings of the 3rd Asian Conference for Information Technology in Asian agricultural information technology & management. Publications of the Japanese Society of Agricultural Informatics, vol. 8, pp. 596–600, 2002.

[23] A. Tharwat, T. Gaber, and A. E. Hassanien, “Cattle identification based on muzzle images using gabor features and svm classifier”, Proceedings of Advanced Machine Learning Technologies and Applications, pp. 236–247, 2014.

[24] S. Mishra, O. S. Tomer, and E. Kalm, “Muzzle dermatoglyphics: a new method to identify bovines”, Asian Livestock, pp. 91–96, 1995.

[25] B. Barry, U. Gonzales-Barron, K. McDonnell, F. Butler, and S. Ward, “Using muzzle pattern recognition as a biometric approach for cattle identification, Trans ASABE, vol. 50, no. 3,pp. 1073–1080, 2007.

[26] H. T. Kim, Y. Ikeda, and H. L. Choi, “The identification of Japanese black cattle by their faces”, Asian Australasian Journal of Animal Sciences, vol. 18, no. 6, pp. 868–872, 2005.

[27] C. Cai, and J. Li, “Cattle face recognition using local binary pattern descriptor”, Proceedings of 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 1–4, 2013.

[28] A. Noviyanto, and A. M. Arymurthy, “Automatic cattle identification based on muzzle photo using speed-up robust features approach”, Proceedings of the 3rd European Conference of Computer Science, ECCS, vol. 110, pp. 114, 2012.

[29] A. I. Awad, H. M. Zawbaa, H. A. Mahmoud, E. H. H. A. Nabi, R. H. Fayed, and A. E. Hassanien, “A robust cattle identification scheme using muzzle print images”, Proceedings of 2013 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 529–534, 2013.

[30] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, “Multimodal deep learning for robust RGB-D object recognition”, In 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 681-687, 2015.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, ..., and A. C. Berg, “Imagenet large scale visual recognition challenge”, International journal of computer vision, vol. 115, no. 3, pp. 211-252, 2015.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks”, In Advances in neural information processing systems, pp. 1097-1105, 2012.

[33] M. Schwarz, H. Schulz, and S. Behnke, “RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features”, In 2015 IEEE international conference on robotics and automation (ICRA), pp. 1329-1335, 2015.

[34] G. Jingqiu, W. Zhihai, G. Ronghua, and W. Huarui, “Cow behavior recognition based on image analysis and activities”, International Journal of Agricultural and Biological Engineering, vol. 10, no. 3, pp. 165-174, 2017.

[35] W. Andrew, C. Greatwood, and T. Burghardt, “Visual localisation and individual identification of Holstein friesian cattle via deep learning”, Proceedings of the IEEE International Conference on Computer Vision, pp. 2850-2859, 2017.

[36] S. Kumar, A. Pandey, K. S. R. Satwik, S. Kumar, S. K. Singh, A. K. Singh, and A. Mohan, “Deep learning framework for recognition of cattle using muzzle point image pattern”, Measurement, vol. 116, pp. 1-17, 2018.

[37] K. P. Risha, K. A. Chempak, and C. S. Sindhu, “Difference of Gaussian on Frame Differenced Image,” International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, vol. 3, special issue 1, pp. 92-95, 2016.

[38] M. S. Norouzzadeh, A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, and J. Clune, “Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning,”, In Proceedings of the National Academy of Sciences, vol. 115, no. 25, pp. E5716-E5725, 2018.

[39] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, “Stacked denoising auto-encoders: learning useful representations in a deep network with a local denoising criterion”, Journal of Machine Learning Research, vol. 11, pp. 3371–3408, 2010.

[40] P. Kamencay, T. Trnovszky, M. Benco, R. Hudec, P. Sykora, and A. Satnik, “Accurate wild animal recognition using PCA, LDA and LBPH”, In 2016 ELEKTRO, pp. 62-67, 2016.

Modeling Network Security: Case Study of Email System

Sabah Al-Fedaghi[1]
Computer Engineering Department
Kuwait University
Kuwait

Hadeel Alnasser[2]
Technical Support Department –
Kuwait Anti-Corruption Authority
Kuwait

***Abstract*—We study operational security in computer network security, including infrastructure, internal processes, resources, information, and physical environment. Current works on developing a security framework focus on a security ontology that contributes to applying common vocabulary, but such an approach does not assist in constructing a foundation for a holistic security methodology. We focus on defining the bounds and creating a representation of a security system by developing a diagrammatic representation (i.e. a model) as a means to describe computer network processes. The model, referred to a thinging machine, is a first step toward developing a security strategy and plan. The general aim is to demonstrate that the *representation* of the security system plays a key role in making thinking visible through conceptual description of the operational environment, a region in which active security operations are undertaken. We apply the proposed model for email security by conceptually describing a real email system.**

Keywords—Network security; conceptual model; diagrammatic representation; email system

I. Introduction

Security typically refers to the state of being secure; that is, being free from danger (the reciprocal of safety). A security system involves deterrence and prevention of threats and is internally motivated as a self-protecting system. It is possible to be in a safe state while danger is present. We can represent this with the notion of threat as a point on a scale between safety and harm that indicates potential danger. Threat also indicates the degree of preparedness required to achieve security.

Preparedness contrasts with vulnerability. Given a certain degree of threat, security is a measure of preparedness for confronting such a threat. The first requirement of preparedness is defining the bounds and creating a representation of the operational environment. The first phase in operational security involves identifying critical assets (e.g., data) and an infrastructure in the operational environment [1]. From the military's perspective [2], the first step in characterizing the operational environment involves identifying the operational area and determining the significant characteristics and level of detail.

According to Tolone et al. [2], "The success of defense and security operations depends on the ability to make sense of the operational environment and to anticipate those factors that influence operations both negatively and positively." Making sense refers to creating situational awareness and employing continuous effort to understand connections, anticipate their trajectories, and act effectively [2].

In this context, this paper imports the sense of operational security in computer network security, including infrastructure, internal processes, resources, information, and physical environment. We focus on defining the bounds and creating a representation by developing a diagrammatic representation (i.e. model) as a means to describe computer network facilities. It is a first step toward developing a security strategy and plan.

A. The Security Problem

Researchers have categorized some security problems as "wicked problems" due to their complexity, intricacy, and intractability [3]. According to Gilmore [4], "Many cyber issues personify a wicked problem. Cyber security can never truly be solved (a completely secure network is a myth)."

Huguet [5] examined the security notion in general, concluding, "Nowadays there are several 'securities' and a number of models from different fields. Despite the importance of the issue, surprisingly, there is no common vocabulary, procedures, definition or model to share knowledge about security." Having a general framework of the security concept, in which to integrate those models and concepts, has advantages, such as shared vocabulary, knowledge, development, or metrics [6-7]. Solms and Solms [7] emphasized that "with the need to implement IT-security measures in almost every environment. Holistic security ontology is still missing. We have to model proper countermeasures capable of protecting the resources." The authors also stressed infrastructure elements such as electronic devices and networks, as well as their relationships.

We note that current works on developing a security framework focus on a security ontology that contributes to applying a common vocabulary, but such an approach does not assist in constructing a foundation for a holistic security methodology: "a holistic formal graphical and textual paradigm for the representation, development, and lifecycle support of complex systems" [8]. Moreover, in security, "the problem lies in the details" [9]. Maintaining security networks with heterogeneous systems, policies, and capabilities quickly became a major task because system administrators were required to maintain detailed descriptions of each host [9].

B. Contributing to the Solution

This paper focuses on developing a modeling language that can be utilized to build a security foundation. The objective is to demonstrate that the representation of the security system plays a key role in making thinking visible, through conceptual description of the security of the operational environment, a region in which active security operations are undertaken. Fig. 1 and 2 show two descriptions of such a theater of operations.

Our proposed representation (model) is constructed in terms of a diagram that serves at the level between "natural communication" and semiformal specification to facilitate understanding among all security participants as a first step toward developing and implementing security policies and implementation plans.

C. Focus

Without loss of generality, in this paper we focus on developing a foundation for email security through conceptually describing what we previously called a theater of operation involving email. Today, email has become the backbone of many professionals' daily activity. Emails are most frequently used in commerce [12]. In everyday life, we rely on email's confidentiality and integrity to exchange data and communication.

According to Landewe [13], email is the primary threat to companies using enterprise platforms, such as Office 365. Email security aims to develop an email technology with a more innovative and multilayered approach to cloud security.

New email security technology involves monitoring user behavior and events, as well as greater access to files, users, and controls. This allows suspicious email to be caught before it reaches an inbox. Using API, it is instead held in a quarantine folder. A copy of the email is run through various technologies (e.g., sandboxing) [13]. The identification of suspicious email is accomplished by performing language and contextual analysis and business email compromise and phishing analysis. Emails with a URL must be handled with link analysis using real-time feeds or by sandboxing the URL [13].

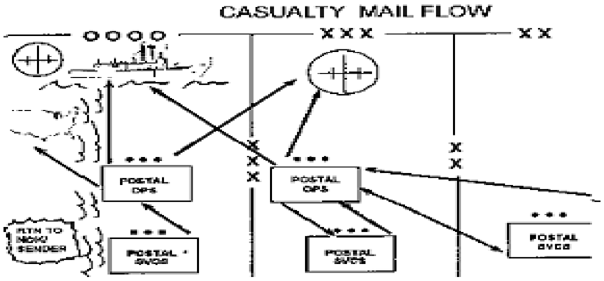


Fig. 1. Mail Theater of Operation (adapted from [10]).

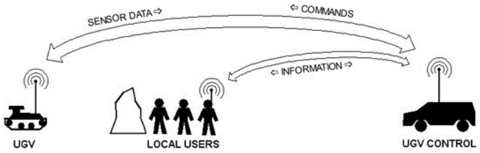


Fig. 2. Sample Description of a Theater of Operation (adapted from [11]).

II. Related Works

The email system architecture is typically introduced when discussing email issues. Two samples of such representations are shown in Fig. 3 and 4. Many types of email network diagrams exist, and UML diagrams are also used (e.g., use-case diagram). Fig. 5 shows a sample of email models. Other works use architectural diagrams that supplement a hardware-oriented network to investigate logical data flow embedded in a system. This approach arrives at a variation of the UML diagrams and includes actual hardware connectivity and logical flow of data (e.g. [14]). In UML, the multiplicity of diagrams is a known problem [8] when what is needed is a single, integrated diagrammatic representation that incorporates function, structure, and behavior.

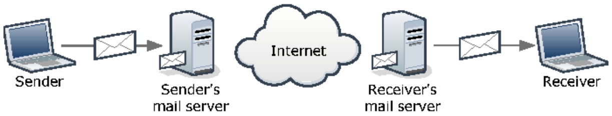


Fig. 3. Email System Architecture (Partially Adapted from [15]).

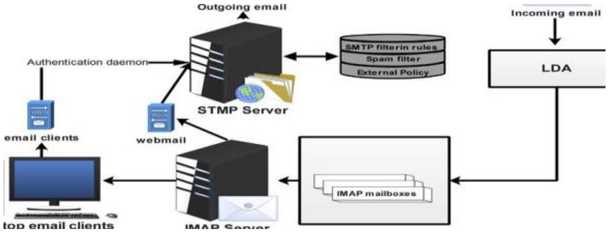


Fig. 4. Email System Architecture (Partially Adapted from [16]).

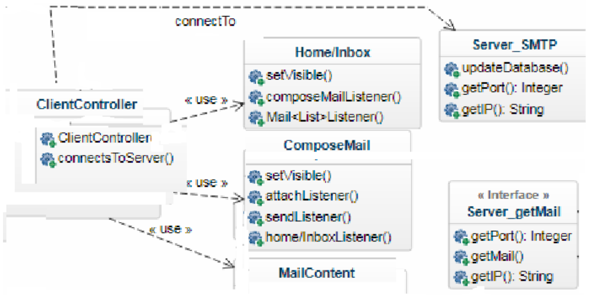


Fig. 5. Sample UML Email System Model (Partially Adapted from [17]).

III. Thinging Machine

According to Wong [18], tackling wicked problems can be achieved by conceptualizing them as systems and breaking them down into "chunks of information" or digestible nodes and their relationships. In describing the model for email operational security, we use the thinging machine (TM) model [19-28] where all operational elements are conceptualized in terms of a single ontological entity, the thimac (thing/machine). As we show, thimacs can represent heterogeneous entities: physical entities (e.g., a server, router, or workstation), software objects (e.g., a program or software system), and other notions (e.g., protocols, flows, or plans).

The term thing (in contrast to objects in object-oriented modeling) indicates an expansive specification of an entity that reflects one mode of an entity's being. Heidegger's [29] notion was that a thing is not a mere abstract object but something that is operated on (created, changed, and transported) and, simultaneously, a process (machine) that subjects other things to its activities (creating, changing, and transporting). According to Heidegger [29], a thing "things"; that is, it gathers, unites, or ties together its constituents, in the same way that a bridge unifies aspects of its environment (e.g., a stream, its banks, and the surrounding landscape) [30].

Building on such a philosophical approach, things are combined with process by viewing them as blocks of single ontological thimacs, which populate a world that is also a thimac (we call it a system). In contrast to the object-oriented paradigm, every part of this world is a thimac, forming a thimac-ing network. A unit of such a universe has dual being as a thing and as a machine. A thing is created, processed, released, transferred, and/or received. A machine creates, processes, releases, transfers, and/or receives things. We will alternate between the terms thimac, thing, and machine according to the context.

The thimac as a {thing, machine} pair designates what simultaneously divides and brings together a thing and a machine (process in the general sense). Every thimac appears in a system either by creation or importation from outside the system. They are the concomitants (required components) of a system and form the current fixity of being for any system that continuously changes from one form (thing/machine) to another. We will use the notion of thimac to model an email system focusing on security aspects.

The terms system and model have been used ubiquitously in engineering [31]. In TM, a system is the overall constellation of thimacs that structures all subthimacs in the problem under consideration. It provides the problem's unifying element through space and time as integral subthimacs, not as the sum of individual subthimacs. Thimacs inside a system are understood not as things with properties but as ensembles of things and machines that constantly interact with each other and with the out-of-system world.

In this complex model, events (a type of thimac that involves time) appear, propagate, and constantly recur in various parts of the system with repeatable occurrences and stable regularities. In its static and dynamic modes, the whole system is a representation (mimesis) of a portion of reality.

Accordingly, a thimac's existence depends on its position in the larger system, as either a thing that flows in other machines or a machine that handles a flow of things (i.e., create, process, release, transfer, and receive things). It brings together and embraces both "thingishness" and "machineness." A thing's flow is conceptualized as an abstract structure that forms an abstract machine called a TM (Fig. 6), in which the elementary processes are called the stages of a TM. In the TM model, we claim that five generic processes of things exist: things can be created, processed, released, transferred, and received. These five processes form the foundation for modeling thimacs. Among the five stages, flow (solid arrow in Fig. 6) signifies conceptual movement from one machine to another or among the stages of a machine. The TM stages can be described as follows.

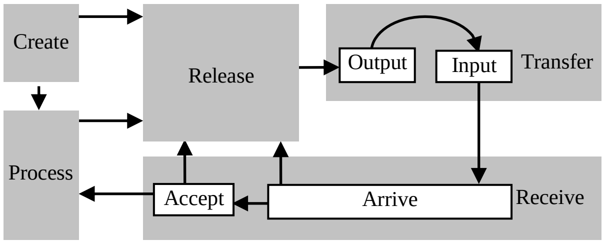


Fig. 6. A Thinging Machine.

Arrival: A thing reaches a new machine.

Acceptance: A thing is permitted to enter the machine. If arriving things are always accepted, then arrival and acceptance can be combined into the receive stage.

Processing (change): A thing undergoes some kind of transformation, without creating a new thing.

Release: A thing is marked as ready to be transferred outside of the machine.

Transference: A thing is transported somewhere outside of the machine.

Creation: A new thing is born (created) in a machine. "Create" resembles "there is."

In addition, the TM model includes memory and triggering (represented as dashed arrows) relations among the processes' stages (machines).

IV. Example of a Thinging Machine

To illustrate TM modeling, in this section we model a communication protocol in widespread use today. One definition of a protocol is a standard description used to define a method of exchanging data over a computer network [32]. In TM, a protocol is viewed as a thimac (machine) that is formed from subthimacs that create, process, release, transfer, and/or receive things (e.g., signals, data, or messages). The protocol is also a thing that can be created and processed (e.g., updated). We apply TM modeling to the well-known simple mail transfer protocol (SMTP).

SMTP is an Internet standard for email transmission [33-34]. It allows for a simple email service and is responsible for moving messages from one email server to another. SMTP includes many standard commands (e.g., EHLO, MAIL FROM, RCPT TO, DATA, and QUIT).

The SMTP protocol and how email works can be explained by tracking the journey of an email message from one person, say, Bob, to another, Alice. Bob composes his message and after inserting Alice's email address, he clicks the "send" button. SMTP governs the communication between Bob's mail server and Alice's mail server [35]. Fig. 7 shows a sequence diagram that models all the events (in the UML sense) involved in this communication, assuming everything works correctly. This diagram is shown in many sources on the Internet (e.g., [35][36][37]).

Fig. 8 shows the corresponding TM model, in which the whole protocol is constructed from the two subthimacs: Bob's mail server and Alice's mail server. In the first machine, the

EHLO message (identifying domain, e.g., Gmail.com) is constructed and sent by Bob's mail server (Circle 1 in Fig. 8). The message flows to Alice's mail server, where it is processed (2) to create a response message (3) of acknowledgement (4), along with the name of the email services that the SMTP server can support (e.g., Yahoo.com). Bob's mail server sends the sender's email address (5; e.g., Bob@gmail.com), which flows to Alice's mail server to trigger an OK message (6) that reaches Bob's mail server. Afterwards, Bob's mail server sends the email address of the recipient (e.g., Alice@yahoo.com) (7) to trigger Alice's mail server to reply with an OK message (8).

At this point, Bob's mail server requests that the data part of the email be sent (9), and upon receiving a ready message from Alice's mail server (10), Bob's mail server starts sending the data (11) line by line. Upon sending the whole message, Alice's mail server sends an acceptance of the message (12). Upon receiving the message, Bob's mail server requests to quit (13) and Alice's mail server sends a signal to close the connection (15), which is closed as the last step (15).

In contrast to the sequence diagram in Fig. 7, the representation in Fig. 8 is based on TM. Although the sequence diagram potentially includes millions of arbitrary actions (e.g., send, identify, terminate, receive, respond, accept, and close), the TM specification repeatedly uses five generic operations. Even though Fig. 8 has the appearance of a complex structure, this complexity is a visual impression that emerges from this repeated application of TM. According to Bishop [38], systems that have a complicated set of interacting parts may actually exhibit relatively simple behavior.

Fig. 8 models the static description of SMTP. To model the dynamic behavior, we use events. An event in TM is a thimac that includes a time machine. For example, Fig. 9 shows the event Sending a line of data. Because a hierarchy of events exists in the SMTP example, we select the 12 events in Fig. 10. Accordingly, Fig. 11 shows the dynamic behavior of the SMTP system.

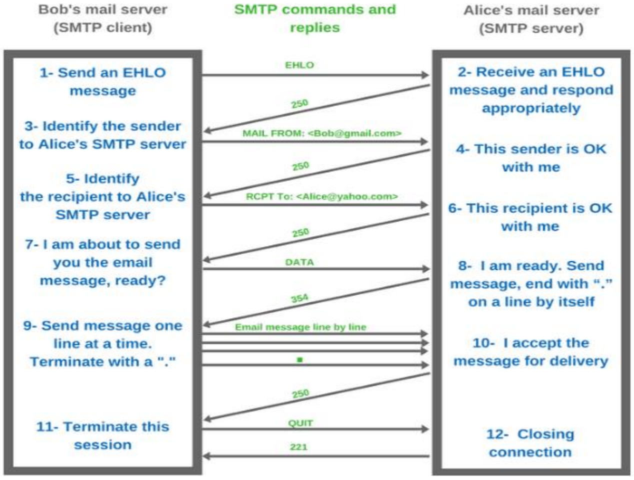

Bob's mail server
(SMTP client)
SMTP commands and replies
Alice's mail server
(SMTP server)
1- Send an EHLO message
EHLO
2- Receive an EHLO message and respond appropriately
250
3- Identify the sender to Alice's SMTP server
MAIL FROM: <Bob@gmail.com>
4- This sender is OK with me
250
5- Identify the recipient to Alice's SMTP server
RCPT To: <Alice@yahoo.com>
6- This recipient is OK with me
250
7- I am about to send you the email message, ready?
DATA
8- I am ready. Send message, end with "." on a line by itself
354
9- Send message one line at a time. Terminate with a "."
Email message line by line
.
10- I accept the message for delivery
250
11- Terminate this session
QUIT
221
12- Closing connection

Fig. 7. Representation of SMTP as a Sequence Diagram.

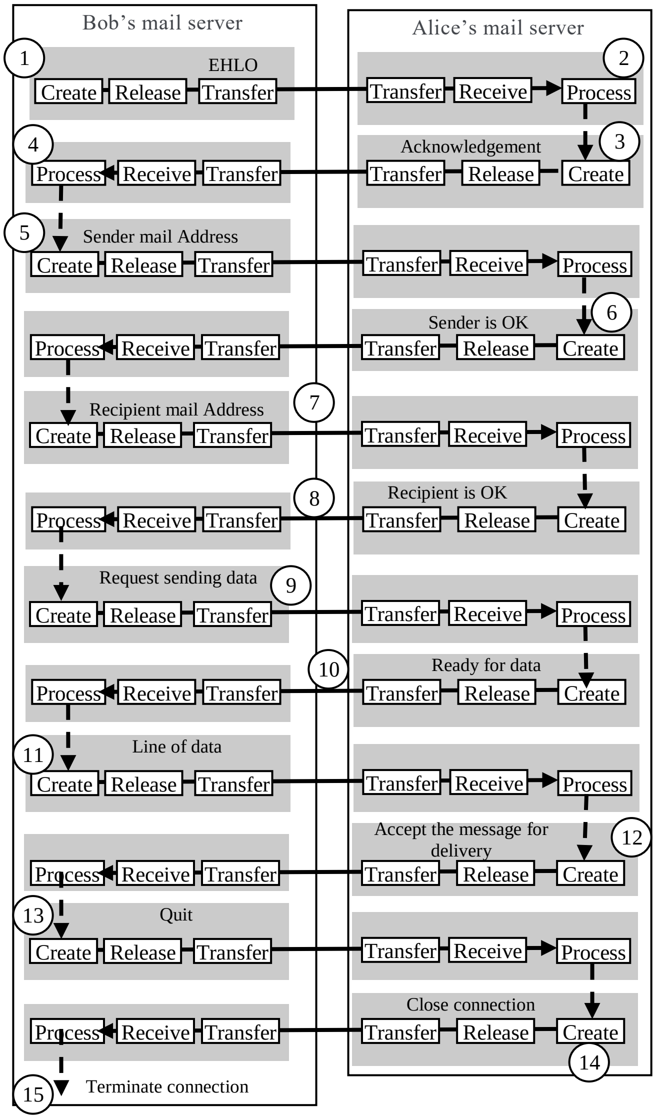


Fig. 8. The SMTP Thinging Machine.

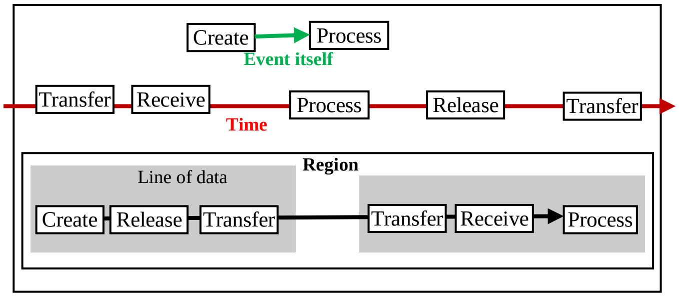


Fig. 9. The Event Sending a Line of Data.

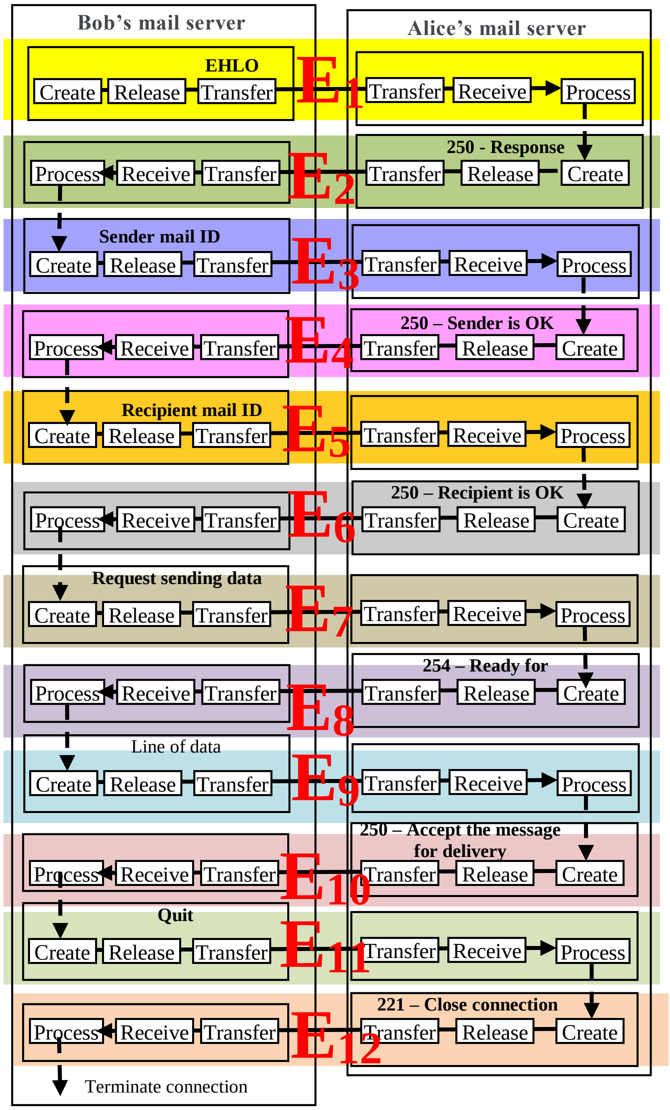


Fig. 10. Events of the SMTP Thinging Machine.

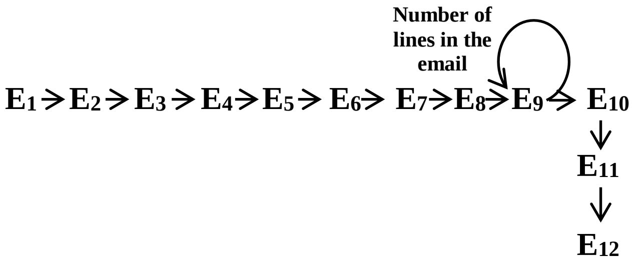


Fig. 11. The Chronology of Events in the SMTP Thinging Machine.

V. Case Study

The map of the email system's security control is essential for understanding, implementation, expansion of the design, training, documentation, and management. The map includes descriptions of each email security device and system that participates in the network. In this case study, we develop such a map for a currently existing government network (workplace of the second author). We limit our mapping to tracking the email throughout the network. This involves diagramming the creation and processing stages of the email's flow from the user workstation to its destination. This includes the following components.

1) User's workstation (e.g., any smart device that can access the organization's email system).

2) Email system: The email system that includes servers that facilitate emailing across the network.

3) Internal firewall: An internal firewall only allows legitimate traffic based on configured policy and rules.

4) Email security gateway: An email security gateway prevents the transmission of emails that violate policy, malware, or transfer of information with malicious intent.

5) External firewall: An external firewall only connects an internal network to an external network and all services or published servers, along with third-party connection (in our case Internet), to separate and secure internal networks and traffic.

6) Domain Name System (DNS) server: A DNS server is a computer server that contains a database of public IP addresses and their associated host names, and in most cases, serves to resolve or translate those names to IP addresses as requested (e.g., www.google.com will be translate to 8.8.8.8). DNS servers run special software and communicate with each other using special protocols.

7) Internet service provider router: A router provides access to the Internet and transfers the traffic from the external firewall to the Internet cloud.

Fig. 12 shows a general picture of the connections among these components of our case study.

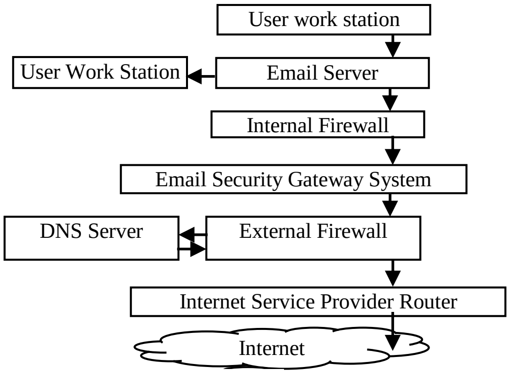


Fig. 12. A General View of the Modelled Network.

VI. TM Email Model

Fig. 13 shows the TM model of the system.

A. *In the user Workstation*

In the figure, the email process starts when the user, on his or her workstation (1), creates an email (2) using the email system (3). This process involves the following.

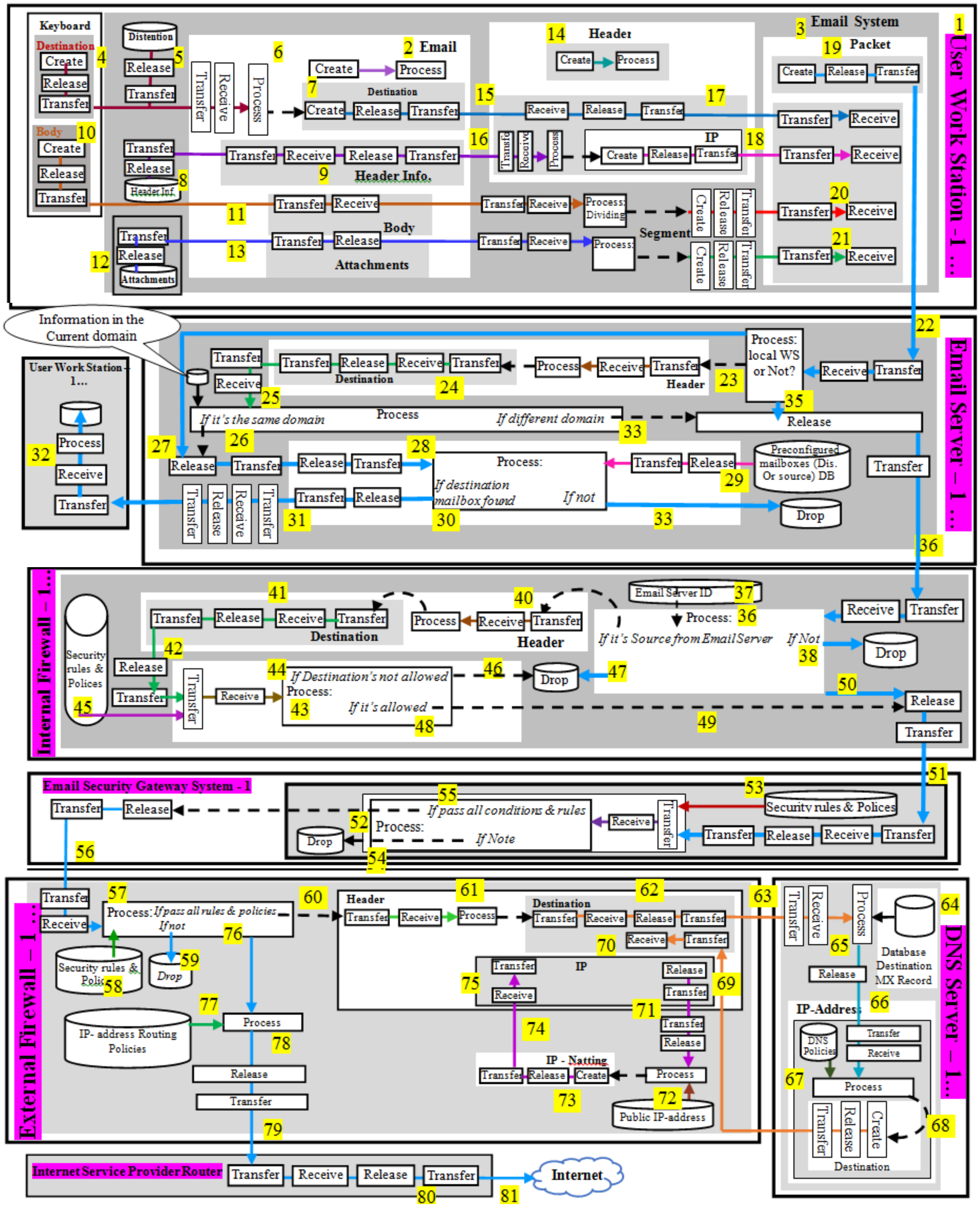


Fig. 13. The TM Model of the Email System.

1) The destination address is either typed in using the keyboard (4) or retrieved from storage (5) to be processed (6) in the email system to create an initial destination format (7).

2) In addition, the header format (8) is retrieved and flows to the email (9).

3) The email body is created (10) and flows to the email (11).

4) If there is attachment (12), then it flows to the email (13).

5) The header is created (14) by combining the destination (15) and initial header information (16) to fill the IP address fields.

6) The email packet is generated (19) by conjoining the header IP address field information (17, 18), body (20), and attachments (21).

B. In the Email Server

The email packet leaves the user's workstation (22) and reaches the email server (known as the exchange server). The packet is processed in the email server to extract the header (23), and then the destination is extracted (24) for comparison with the information in the current domain.

1) If the destination has the same current information domain (26), then the packet flows (27) to be processed (28) for comparison with preconfigured mailboxes (29).

- If a destination mail box is found (30), then the packet flows (31) to the user's work station (32).
- If the destination mail box is not found (33), then the packet is dropped.

2) If the destination does not have the same current information domain (34), then the packet flows (35) to the internal firewall (36).

C. In the Internal Firewall

In the internal firewall, the packet is checked (37) for an email server ID (38):

1) If its source is not from the current email server (39), then it is dropped (40).

2) If its source is from the current email server, then the header is extracted (41) and the destination in the header is extracted (42). Then the destination flows (43) to be processed (44).

-If the destination is not permitted (45) by the security rules and polices (46), then the packet is dropped (47, 48).

-If the destination is permitted (49) by the security rules and polices (46), then the packet flows to the email security gateway system (50, 51, and 52).

D. In the Email Security Gateway System

In the email security gateway system, the packet is compared (53) with the security rules and polices (54).

1) (i) If the packet does not satisfy all polices and rules, then it is dropped (55).

2) (ii) If the packet is passes all polices and rules (56) then it's flow to the External Firewall (56).

E. In the External Firewall

The packet is processed (57) in the external firewall.

1) If the packet does not satisfy the security rules and polices (58), then it is dropped (59).

2) If the packet satisfies all polices and rules (60), then the header information (61) and the destination (62) are extracted from the header. The destination flows to the DNS server (63) to be compared with the stored DNS database records (64) to select the related destination MX record (65). The MX record, known as the mail exchanger record, specifies the mail server responsible for accepting email messages on behalf of a domain name. It is a resource recorded in the DNS. This record flows to be processed again (66 and 67) with the DNS polices to create the destination IP address (68) that flows to the header (69 and 70).

The IP address of the header flows to be processed (71) with the public IP address (72) to be processed again to create the natted public IP address that is used for the routing polices (73), which flows again to the header (74 and 75).

Once the header is updated, the packet is released (76) to be processed (77) with the IP address routing polices (78) to learn its next destination, then it travels (79) to the Internet service provider router.

F. In the Internet Service Provider Router

The email packet leaves the external firewall and is transferred (80) to the cloud (81).

VII. TM Dynamic Model

Note that the static email model structure is formed from the flows among thimacs in the system. It also includes the network of thimacs belonging to the system. The resultant conceptual model is a representation of structure in the email system. An email has a class (e.g., object-oriented) form with header, data, and attachment attributes (object-oriented terminology). These are filled by flows to produce an object (object-oriented terminology). The behavior is yet to be defined when we incorporate events in the static model. Note the two forms of a thimac. For example, the packet is a machine that is fed addresses, pieces of data, and attachments. As soon as it is loaded, it flows as a thing to the email server.

A thimac is activated by elevating it to a time thimac (e.g., a subdiagram of Fig. 13 becomes the region of an event). We can develop the dynamic model of the email system as we did before, for the SMTP protocol. However, in consideration of space, we only identify events in the user workstation and the pre-email server as shown in Fig. 14. Thus, what appeared in Fig. 13 as the transmission of one packet, in the dynamic model with its chronology of events (Fig. 15), we see as a repeated generation of packets until all data and attachments have been sent.

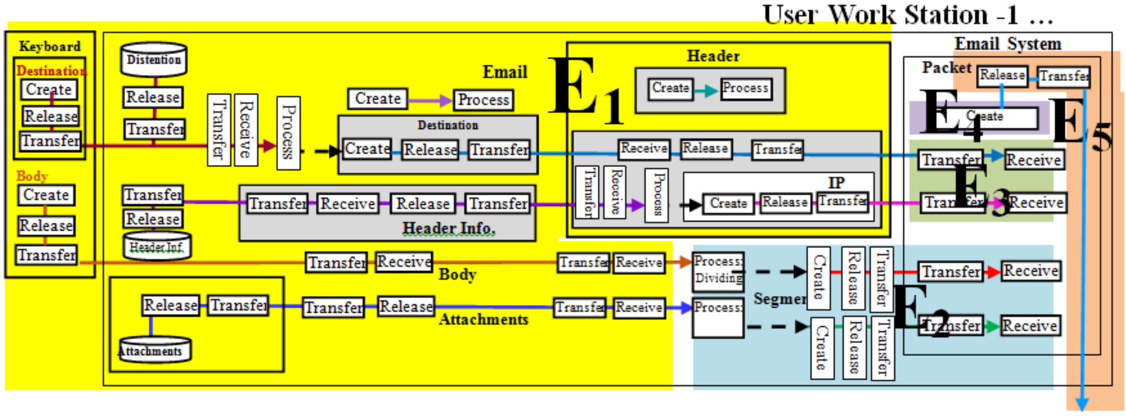


Fig. 14. Selected Events in the user Workstation.

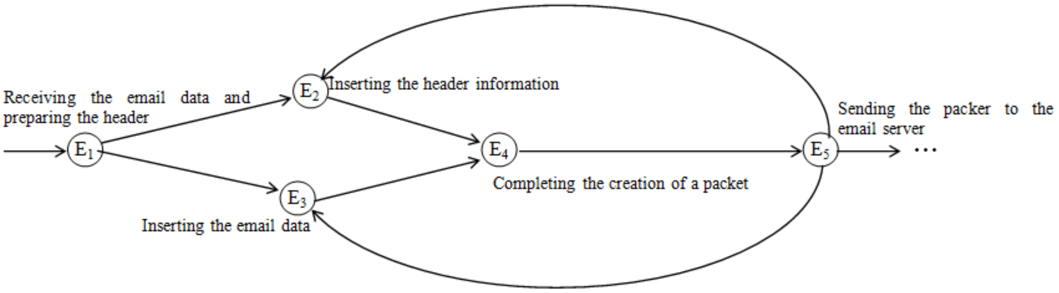

Receiving the email data and preparing the header
E1
E2
Inserting the header information
E3
Inserting the email data
E4
Completing the creation of a packet
E5
Sending the packer to the email server
…

Fig. 15. Chronology of Events in the user Workstation.

VIII. TM Security Modeling

In the TM network model, we can separate separate the network security system from the functional system. Each security thimac (hardware and software) can be described by tracking the packet flow. The ad hoc diagrams and flowcharts currently in use (e.g. see Fig. 16 and 17) are “crude” representations that are not as systematic as the TM model, which involves only five primitive operations. In addition, these current representations do not model dynamic aspects in the same diagram, e.g., events of firewall-1 as shown in Fig. 18.

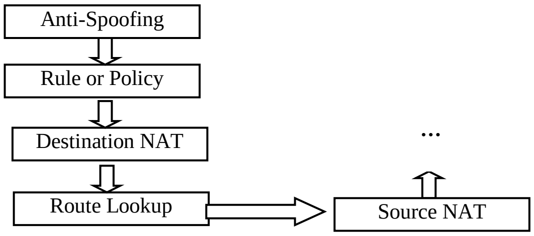


Fig. 16. Packet flow Check Point Firewall (Partially Adapted from [39]).

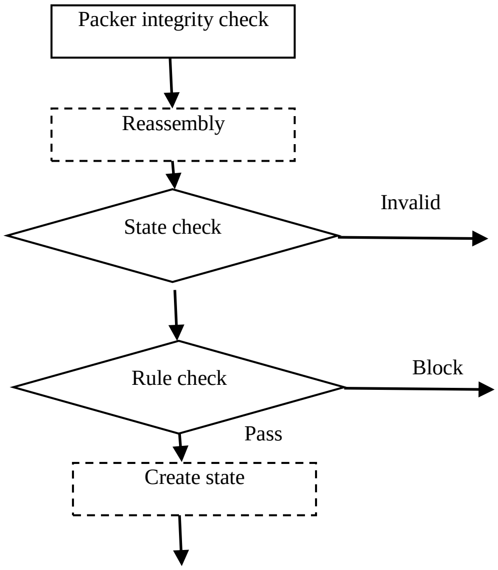


Fig. 17. Packet Processing in Firewall (Partially Adapted from [40]).

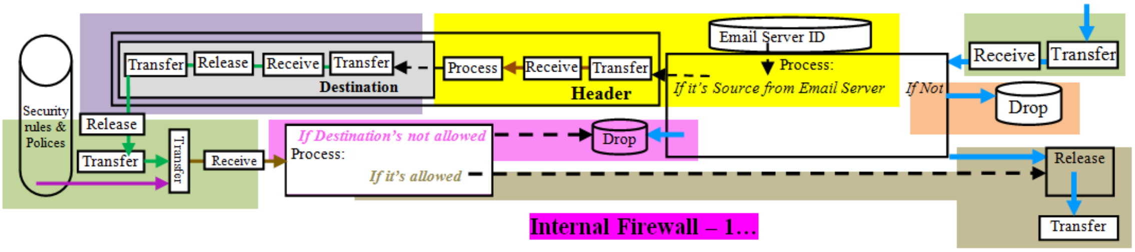


Fig. 18. Events in Firewall-1 (see Fig. 13).

IX. Conclusion

In this paper, we claim that the task of building a security system requires construction of a diagrammatic representation of the operational environment where protected assets reside. We criticized current modeling languages (e.g., UML and ad hoc diagrams) as languages that either include large heterogeneous notions or nonsystematic symbols (wall, cloud, screen, human figure, etc.). Instead, we have proposed using TM modeling and applied TM language to an actual email system by tracking security aspects as an email flows through various system components. The resultant map can be used as a foundation for the activities of an email security officer, just as a network diagram is a tool for a network engineer. Future work will apply the TM model to different types of networks.

References

[1] J. Andress, “Operations security,” in The Basics of Information Security (2nd ed.). Elsevier, 2014.

[2] W. J. Tolone, X.. Wang and W. Ribarsky, Making Sense of the Operational Environment through Interactive, Exploratory Visual Analysis, NATO OTAN unclassified paper.

[3] P. Williams, Security Challenges as Wicked Problems, Stratfor, Jun 1, 2018.

[4] K. Gilmore, “Cyber security: A wicked problem,” Air Force Information Technology & Cyberpower Conference, Aug 29-31, 2016, Montgomery, AL, 2016.

[5] M. Colobran Huguet, “A general-purpose security framework,” Ph.D. thesis, Universitat Autònoma De Barcelona, September 2015.

[6] S. Fenz and E. Weippl. “Ontology based IT-security planning,” in Proceedings of the 12th Pacific Rim International Symposium on Dependable Computing, pp. 389–390. IEEE Computer Society, University of California, Riverside, 2006.

[7] B. von Solms and R. von Solms, “From information security to business security?” Computers & Security, Vol. 24, No. 4, pp. 271–273, 2005.

[8] D. Dori, Object-Process Methodology - A Holistic Systems Paradigm. Berlin: Springer Verlag, 2002.

[9] M. Carvalho, M. Rebeschini, J. Horsley, N. Suri, T. Cowin, M. Breedy, “MAST: Intelligent roaming guards for network and host security,” Scientia, Estudos Interdisciplinares em Computação, Vol. 16, No. 2, pp. 125-138, December 2005.

[10] GlobalSecurity.com, Postal operations management, Chapter 6, 2020.

[11] H. Sallum, “Development and implementation of a high-level command system and compact user interface for non-holonomic robots,” M. S. thesis, Massachusetts Institute of Technology, May 2005.

[12] S. Choudhary, “E-mail security: Issues and solutions,” International Journal of Computer Information Systems, Vol. 7, No.4, 42-46, 2013.

[13] J. Witts, “M. Landewe interview,” Expert Insights, Nov 20, 2019.

[14] T. Olzak, “A practical approach to threat modeling,” Erudio Security, LLC, 2006.

[15] A. Z. Adamov, “Internet technologies in depth. The technique of spam recognition based on header investigating,” 5th International Conference on Application of Information and Communication Technologies (AICT),1 - 5, Azerbaijan, Baku, 12-14 October 2011.

[16] E. Gbenga Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, “Machine learning for email spam filtering: Review, approaches and open research problems,” Heliyon Vol. 5, No. 6, 2019.

[17] Oshinsingh, “Email system, UML model,” GenMyModel, September 24, 2014.

[18] E. Wong, “Wicked problems: 5 steps to help you tackle wicked problems by combining systems thinking with agile methodology,” The Interaction Design Foundation, 2019.

[19] S. Al-Fedaghi and A. J. Al-Fadhli, “Thinging-oriented modeling of unmanned aerial vehicles,” International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 11, November, 2019.

[20] S. Al-Fedaghi and Y. Atiyah, “Tracking systems as thinging machine: A case study of a service company,” International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 9, No. 10, pp. 110-119, 2018.

[21] S. Al-Fedaghi and M. BehBehani, “Thinging machine applied to information leakage,” International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 9, No. 9, pp. 101-110, 2018.

[22] S. Al-Fedaghi and M. Al-Otaibi, “Conceptual modeling of a procurement process: Case study of RFP for public key infrastructure,” International Journal of Advanced Computer Science and Applications (IJACSA) – Vol. 9, No. 1, January 2018.

[23] S. Al-Fedaghi and N. Al-Huwais, “Conceptual modeling of inventory management processes as a thinging machine,” International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 9, No. 11, pp. 434-443, November 2018.

[24] S. Al-Fedaghi, “Thinging as a way of modeling in poiesis: Applications in software engineering,” International Journal of Computer Science and Information Security (IJCSIS), Vol. 17, No. 11, November 2019.

[25] S. Al-Fedaghi, “Thing/Machine-s (thimacs) applied to structural description in software engineering,” International Journal of Computer Science and Information Security (IJCSIS), Vol. 17, No. 8, August 2019.

[26] S. Al-Fedaghi, “Five generic processes for behaviour description in software engineering,” International Journal of Computer Science and Information Security (IJCSIS), Vol. 17, No. 7, pp. 120-131, July 2019.

[27] S. Al-Fedaghi, “Modeling events and events of events in software engineering,” (IJCSIS) International Journal of Computer Science and Information Security, Vol. 18, No. 1, 2020.

[28] S. Al-Fedaghi, “Thing/machine-s (thimacs) applied to structural description in software engineering,” International Journal of Computer Science and Information Security (IJCSIS), Vol. 17, No. 8, August 2019.

[29] M. Heidegger, “The thing,” in Poetry, Language, Thought, A. Hofstadter, Trans. New York: Harper & Row, 1975, pp. 161–184.

[30] M. Heidegger, Being and Time, J. Macquarrie and E. Robinson, Trans. London: SCM Press, 1962.

[31] D. Hestenes, “Notes for a modeling theory of science, cognition and instruction,” in E. van den Berg, A. Ellermeijer & O. Slooten, Eds., Modelling in Physics and Physics Education. U. Amsterdam, 2008.

[32] N. Emberton, “Protocol,” Computer Hope, Oct. 7, 2019.

[33] J. B. Postel, “Simple mail transfer protocol,” Information Sciences Institute, University of Southern California, August 1982.

[34] J. Klensin, “Simple mail transfer protocol,” Network Working Group, October 2008.

[35] K. Elghamrawy, “SMTP protocol explained (How email works),” Afternerd Blog, 2017-2019.

[36] C. M. Kozierok, The TCP/IP Guide, September 20, 2005.

[37] TeleMessage, “How to know if the mail server on the other side received an email,” TeleMessage Site, 1999-2020.

[38] R. C. Bishop, “Metaphysical and epistemological issues in complex systems,” in Handbook of the of Science Philosophy, Vol. 10, edited by Cliff Hooker, North Holland, 2011.

[39] F. Ali, “Checkpoint firewall packet flow,” Network Engineer, April 28, 2019.

[40] Oracl Documentation Center, Securing the network in Oracle Solaris 11.4, January 2019.

Usability Study of Smart Phone Messaging for Elderly and Low-literate Users

Rajibul Anam[1]
Smartphone App Research Division
Anam Research & Development Solutions
Dhaka, Bangladesh

Abdelouahab Abid[2]
Faculty of Computer and Information Systems
Islamic University of Madinah
Madinah, Saudi Arabia

***Abstract*—Smartphones are electronic devices that people can carry around and install/add compatible third-party Apps to expend their functionality. Smartphones are mainly developed for calling and messaging purposes. All applications' interfaces are designed for the current trends. Therefore, Senior Citizen and Low-literate users face difficulties to use smartphones due to the perceived complicated interface and functionality. This paper analyzes Senior Citizen and Low-literate user's requirements to read and write messages from users "memory load", "navigation consistency", "consistency and standard", and "touch screen finger-based tapping" perspective. Then a framework based on "visual representation", "navigation" and "miss click avoidance" is developed. A comparison between the proposed application and other messaging applications is provided. This research work focused on the Senior Citizen and Low-literate users to improve their user experience of the smartphone messaging application.**

***Keywords*—Smartphone interface; smartphone messaging; visual color; adaptation**

I. Introduction

The smartphone is a device that runs by its own independent operating system, supports touchscreen functionality, very small in size, has calling functionality, can install third-party applications and can carry anywhere [1]. The cell phone was developed only for calling purpose but due to the smartphone with modern technology, it becomes very popular to use and can run much application simultaneously [2]. The smartphone applications are developing very rapidly, with these applications users can make phone calls, reminders, planning, internet calls, online message, sending emails, banking operations, games, etc. [3]. Recently the wearable technology becomes very popular with the young generation [4]. Everyday a variety types of the smartphone are introducing in the market by different companies and each manufacturing companies has its own different functionalities [5]. The user interface should follow a pattern to serve the diversity of the application. An application should be accessed by as many people as possible with the maximum range of environment [6]. Mobile user interface design is not as easy as a desktop app interface design. The functionalities of some mobile devices are very easy for some users and at the same time, it is very hard to understand [7].

Around the globe, five billion people are using mobile services [8]. Among these users, most of them are young and have advanced knowledge of technology. According to the World Health Organization, the world's population will cross two billion who's age is sixty over by 2020 [8], [9]. Among them, those who are over eighty years having movement and motor problems, moreover some users having motor problems after fifty years [10-11]. Most of these Elderly and Low-literate users do not use the smartphone. The main barrier is the usability, short term memory and technology adaptation phobia [12-15]. The usability refers to user satisfaction, how effective and efficient way to complete the task with the given interface with less error [5], [16]. Because smartphone is equipped with touchscreen instead of button keypad [6], limited screen size 3.2" up to 12.9" [17], poorly design for the senior adults which is very hard to adapt [18-19], icons are very small in size [20-21], text entry which refers to incorrect input [18], [22], inconsistence navigation to go next page or previous page [7], [23], and hand/motor function is slow compared to the younger user which required touch operation [24-25]. The mobile manufacturing companies are developing mobile layout according to the trend to do commercialized their product. No one is considering the Elderly Citizens, what interface will give them comfort to use the device [26]. As a result, there is a big number of people in our society who are Elderly and try to avoid using smartphones [27-29]. Some researchers think that the next generation of Elderly citizens will be different from the Current Generation of Elderly citizens. This Generation is habitual with the present system and when they become Elderly Citizen will be easy for them to use the smartphone [30]. But still, they will have to face cognitive and motor function issues [24].

Most of the Elderly Citizen uses the Button Phone, which is called non-smartphone [28]. Because of the advanced technology most smartphone users use third-party apps to communicate with each other [2]. The Elderly and Low-literate users use the smartphone only to communicate with their grandchildren and family members through third-party apps including Facebook, WhatsApp, Tumblr, Instagram, WeChat and Twitter [17], [31-32]. There is a big gap in usability between non-smartphones and smartphones. This paper proposes a system where tries to resolve the gap between using the non-smartphone to the smartphone.

This paper is arranged as follows. Section 2 provides a brief review of other Smart Phone System, which is made for Elderly and Low-literate users, Section 3 describes the proposed solution, Section 4 details discussion, and results of the proposed methods and finally, the conclusion is in Section 5.

II. Related Works

Usability is based on Learnability, Efficiency, Memorability, Satisfaction, and Errors [5]. The Heuristic Evaluation is a set of usability principles, which has evaluated by the experts and collects a list of usability problems [16].

A. Usability

Smart Phone Accessibility Testing and Evaluation focus on the WCAG 2.0 best practice. Author in [17] research focus on the app Interface labels, fonts, colors, and buttons size, moreover on the consistency of the navigation. The Usability study focuses on five social networking apps including Facebook, Tumblr, WhatsApp, Twitter and Instagram. The research found that most of the apps do not follow the color guideline and very poor readability. The Multimedia contents navigations are not in the proper place to interact with. The usability test shows that non-experienced Elderly participants face problems to navigate and finish the specific task on time.

Mobile Health System is a standalone application for the elder and low literate users. Author in [33] design interface (UX) for Glucose Measurement application prototype followed the guideline of UU principle which is under mobile health guidelines and the development tool named Balsamiq and also did usability study on Elderly people who are over fifty years older. The usability study was done by ten specific tasks with Control and Experimental Group of People. The Experimental Group performance was better than the Control Group. This study focus on user preference design, options they prefer to use. This study does not focus on multimedia content.

Mobile user's behavior and UX design make the user's satisfaction level high. Author in [5] followed the standard life cycle of usability testing. The usability test conducted by six types of tasks which is related to the multimedia contents and each task contains six types of instructions with five participants. The tasks asked the Elderly users to use a web browser, calendar and image gallery. After finished the testing the participants give feedback that application navigation and options are not clearly visible on the screen.

Multi-Touch Mobile App Interface is a vital issue for the Elderly and Low-literate users. Author in [12] put focus on the Age-Related Changes, Suitable Input at different model devices, Multitouch Interface users and Designing of Multitouch Interfaces for Elderly. After analyzing these issues, create an initial design guideline for the app development to Target the Interface Design, Use of Graphics, Navigation and Errors, Content Layout Design, User Cognitive Design, Audio, Text Design, User Feedback and Support, Multi-Touch Interaction and Interface Testing. The design guideline was created with structure, in detail and very comprehensive which overcome the age-related changes which might create impact the usability of the multi-touch interface.

Aging is a process of losing physical and cognitive abilities over time. Author in [34] worked on the human coming age, touch interfaces and developed a framework for how touch screen suites Elderly. The researchers collect Elderly user's Smartphone Experience data, touch the interface action activities, look and feel about the buttons, functions they use most. The users feel comfort interface like TV remote control type buttons. At the experiment time found human errors to register the actions like button press pointing position was wrong. They proposed a framework consist of four factors tangible manipulation, spatial interaction, embodied facilitation and expressive representation. The big factor of designing the apps on the assumption of a group, they are sometimes not able to see what the small screen is requesting, which button needs to press and how long or how hard it needs to press the area of the button.

B. Heuristic Evaluation

There are many User Interface has been developed for the Elderly and Low-literate users but most of the designs are rejected by the Usability Study. Author in [1] tried to identify the major usability problems while interacting with the mobile. To improve the User Interface, they proposed a heuristic framework named SMASH which is consists of twelve usability cases. There were five experts involved in this evaluation. The study shows that twenty-seven heuristic violations encountered. The violations were to “Minimized the user's memory load” and “match between system and real-world”. The usability problems are classified into four groups as appearance, language, dialogue and information. This study shows some critical points which need to improve for better Elderly and Low-literate user experience.

Author in [16] Heuristic Evaluation is organized in two stages, one is validation regarding traditional heuristics and another one is validation regarding outcomes from the test with real users. The test was run by the Aptor Software, the application is developed for the Brazilian Elderly people. The Heuristic results show that fourteen factors are not following the rules and nine usability problems countered at the testing period. So, the experts provide a guideline to the developers to make the app more interactive for the Elderly.

Mobile Learning for the Elderly and Low-literate users’ is vital to the present days. Author in [8] developed mobile learning applications to overcome the accessibility issues in order to maximize the usage of the smartphone by Elderly. At the testing period, they found certain accessibility issues for the Elderly. The study was conducted with thirty-two scenarios in a controlled environment. The results come up with certain errors and the experts create a guideline for the smartphone application developers which will help them to makes the application interactive for the Elderly.

Six Smartphone Launchers were introduced for the Elderly and did Heuristic evaluation [11]. The evaluation was classified into three categories firstly look and feel, secondly interaction and lastly functionality. Look and Feel followed by twenty-one criteria, interaction procedure followed by fifteen criteria and functionality followed by thirteen criteria. The evaluation result shows that thirty-nine percent problem list out and sixty-one percent with no problem. The problem listed from three classifications. The results help the future smartphone launcher developers to make an interactive app for the Elderly.

III. Understanding the Issues and Collect Requirements

A. *Types of Senior Citizen*

Senior Citizen or Elderly is defined those are aged sixty and above [18]. But still, there are different types of Elderly exist in our society according to their ability.

1) Fit older people: Elderly who are fit to do most of the work and physical functionalities what they want to do but activities are different from while they were younger [6], [35].

2) Frail older people: Elderly who are fit with some disabilities and having some problem to do regular physical activities [6], [35].

3) Disable person: Who Grow Older Elderly having long term physical disabilities and affected by the aging moreover dependent on their own physical functionalities [6], [35].

The Elderly face problem to understand vision, hearing, dexterities, understanding the menu and navigation links, social contact and mobile application infrastructure [1], [6].

B. *User's Memory Load*

Aging creates memory problems which may start from mid-level of age, memory loss makes human forgetting names, phone numbers, moving objects from one place to another [9]. The smartphone user interface should provide objects which are visible (make elements easy to read), options (easy to accessibility and recognition) and actions (make clickable items easy to target and hit) to prevent irrelevant information from user memory [1], [16], [23]. The menu path should be easy to remember [14].

C. *Navigation Consistency*

Navigation is very important to complete a task effectively and efficiently [1]. The navigation should be consistent and straightforward [7], [23]. Navigation becomes very important when the system is used by the Elderly. Most of the time Elderly open apps and face problem to go back to previous state or face problem to close the app [1], [17], [11], [36], one application put the back button at the top of the page and another app put the back button at the bottom of the page [1]**,** [11]. The back button should be the same place on every screen [17], [31]. The back button should navigate back one level each time [7]. Menu navigation is another factor that makes the Elderly and Low-literate users confused most of the time [6]. The menu is implemented at top of the page by one app and another app puts the menu at the left or right of the page [6]. The app menu should be consistent so the user can access it very easily.

D. *Consistency and Standard*

The smartphone user interface should be comforting to do a task in a familiar, standard and consistent manner [1]. The labels and buttons are not marked correctly and not large enough to initiate or interact with [16-17], [31], font size and spacing is not consistence and appropriate [9], [11], [13-14], [16-17], icon size is small to read and icon symbol is not familiar [9], [12], [15-17], [20], objects color reflection and contrast is not visible (visual acuity) and does not fulfill the readability for the Elderly [6], [11-13], [17], [37]. Displaying content information of the apps sometimes creates confusion, there are different methods to display contents like vertical list, thumbnail list, fisheye list, carousel, grid and film stripe [6], [11].

E. *Touch Screen Finger based Tapping*

Motor, cognitive, and physical abilities make a person perceive and process information [6]. Button design is one of the most common elements on all platforms. There are two types of touch screen, first one is Resistive Touch and second is Capacitive Touch Screen, there is no difference at user end functional difference, these two are hardware issues [38]. There are many types of buttons in mobile applications like action button, radio button, list button, text button, toggle button, icon action button and floating action button [6]. Small size button requires standard human motor control, different timing of contact on buttons may alter the actions. Moreover, it is more complex when it is about to design a button or icon to serve a specific purpose [6], [12], [34]. There are two types of touch interaction supported by smartphones, the first one is a single touch and the second one is multi-touch [39]. Smartphone usability measurement is different because touch screen allows direct interaction using fingers to register a button press by objects (button text and icon), moreover smaller object leads to finger occlusion and wrong point of the press [6], [11], [14], [40]. The Elderly always target to press the target button but due to motor problem, they press outside of the target object and press the target object more or less time it needs to register the request [7], [14], [18], [25], [40-41].

IV. Design and Development

There are two types of design issue named functional and nonfunctional, which can improve the usability for the Elderly and Low-literate users.

A. *Functional*

The functional design specifies a function that the system component must be able to do in a discrete manner. In other words, functionality is describing the behavior of the system [7].

B. *Nonfunctional*

The nonfunctional design is related to usability, performance, acceptance, reliability, quality, effectiveness. In other words, nonfunctionally is an overall property of the system of a particular aspect which is not a specific function [7].

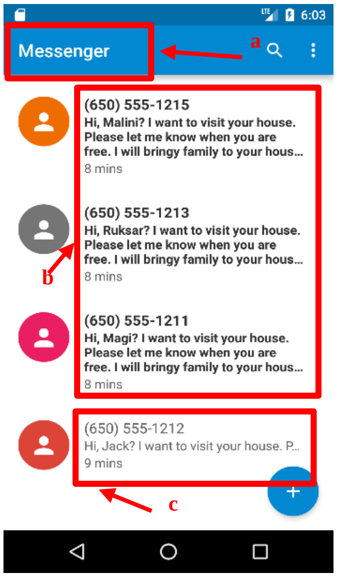
C. *Visual Representation*

Visual representation is a very important factor for touch screen devices, firstly the vision, which gives a clear message to the user, secondly the visual buttons input boundaries which work as button press and finally the text input boundaries should be large enough to give input [7]. The page title fonts size needs to be big and background color should be highlighted with green or blue color [15-17], [31] and same time needs to add an illustration to represent the meaning so the Low-literate users will get a clear idea of the page.

1) Title and illustration of the page: Every page contains its basic information like the name of the page, so the user can easily trace down where or which place they are in. Fig. 1 shows the Google Messages [42] Inbox page, this stage user does not know which page they are now. Fig. 1(a) shows only the Application name, if the user is Elderly or Low-literate user then might face problem to understand the purpose of the page, (b) shows the unread received messages and (c) shows the read received messages. The unread messages are dark black in color and read messages are dim black in color. For the Elderly or Low-literate users, this design is not appropriate. Fig. 2(a) clearly represents the INBOX by illustration and name. The title background color is Green and the font color is white, which will give a clear view of the content. The inbox title gives the idea of the page and the illustration gives a clear idea of the page to the Elderly and Low-literate users, (b) Status notify the users to weather the user previously read the message or not. If the user opens it once the status will be Blue in color and not opened then the status color will be Red and text will be Not Read at white background, (c) provides the message received time. It provides message received hourly and day information exactly bottom of the message summary, finally (d) provides the message summary, maximum three lines will be shown here, for a clear view of the content, the background color used as white and font color Dark Green. Fig. 3 shows the Inbox Display algorithm, where line 1-3 uses the messageobject to check and update the status of the message status, line 4 get the mobile screensize and totalmessagesize value. The screensize get the mobile screen dimension value (different mobile different screen sizes) and totalmessagesize count the number of characters in the message body. Line 6 converts the screensize and totalmessagesize to pixels and line 7 assign the summerymessage as part of the text message which will fit within 3 lines according to the mobile screen dimensions.

2) Arrange distance between objects: Elderly and Low-literate users are not familiar with the technology. Fig. 4 shows the Google New Message page, where (a) shows the gap between two objects, (b) input recipient number and (c) input text message. The input recipient number (b) and (c) input text message height is very small and will create human error to touch the specific points. Moreover, the distance between two objects (b) and (c) is long, which might make the users confused.

The proposed solution Fig. 5 shows the minimal distance from one object to another according to the user prospect, (a) shows the distance between the recipient name and message body, (b) shows the distance between two messages and in the middle there is a thin line which makes them separate. The status notification (c) shows message status with reading, not read and time of receiving. Fig. 6 shows the New Message page, where (a) shows the minimum distance from the recipient's name and text message input boxes, (b) shows the thin borderline outside of the input objects. So, the users can easily understand where they have to press their fingers to register an action.



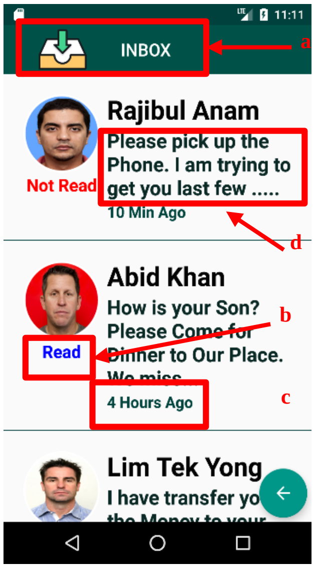
(650) 555-1215
Hi, Malini? I want to visit your house. Please let me know when you are free. I will bringy family to your hous...
8 mins

(650) 555-1213
Hi, Ruksar? I want to visit your house. Please let me know when you are free. I will bringy family to your hous...
8 mins

(650) 555-1211
Hi, Magi? I want to visit your house. Please let me know when you are free. I will bringy family to your hous...
8 mins

(650) 555-1212
Hi, Jack? I want to visit your house. P...
9 mins

Fig. 1. Google Message Inbox.



Rajibul Anam
Please pick up the Phone. I am trying to get you last few .....
Not Read
10 Min Ago

Abid Khan
How is your Son? Please Come for Dinner to Our Place. We miss...
Read
4 Hours Ago

Lim Tek Yong
I have transfer yo
the Money to your

Fig. 2. Proposed Message Inbox.

```
Input:
        messageobject is the total message contents;
Output:
        summerymessage is the total message contents;
Variables:
        screensize is the mobile screen size;

InboxDisplay(messageobject)
1.      If messageobject is open once Then
2.        messageobject status is Read;
3.      Else messageobject status is Not Read;
4.      Calculate screensize and totalmessagesize
5.      Do
6.        Get screensize pixels and threeline pixels;
7.        Assign summerymessage content maximum three line of
        textmessage;
8.      End
```

Fig. 3. Inbox Display Method.

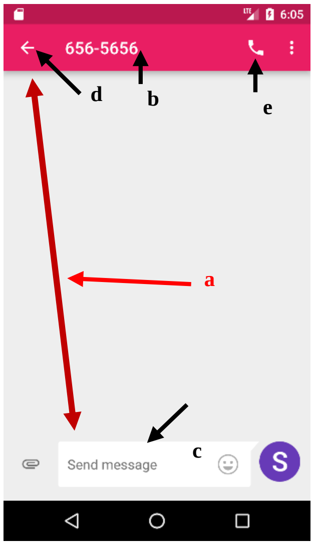
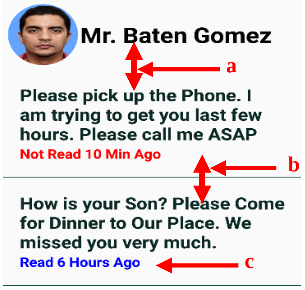


Fig. 4. Google Write New Message.



Mr. Baten Gomez

Please pick up the Phone. I am trying to get you last few hours. Please call me ASAP

Not Read 10 Min Ago

How is your Son? Please Come for Dinner to Our Place. We missed you very much.

Read 6 Hours Ago

Fig. 5. Distance between Objects.

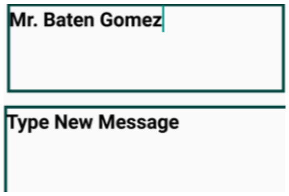


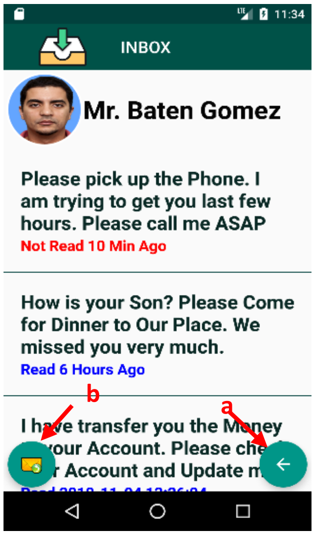
Fig. 6. Border at Input Objects.

D. Navigation

The navigation and menu are major factors to operate any application moreover, the Elderly and Low-literate users face problems to use smartphone regular menu because of their mental model, most of them are familiar with the linear menu style instead of the hierarchical menu [7]. The navigation links visual text and color needs to be very clear to the users after click which page will appear to the user and how to return back to the main page these issues need to be very clear in the navigation.

1) Floating navigation: Fig. 1 is the Inbox screenshots, there is no back button and Fig. 4(d) has back navigation, from these both pages it is visible that back navigation is not consistent. There is back navigation at all the pages, Fig. 7, 8(a) shows the back-navigation button which is consistence at the right bottom of the page. The proposed system Fig. 7, 8(b) design in a manner so the most important navigation action stands the left bottom of each page. Fig. 7 is the inbox page, where reply navigation is the most important action (b), same way Fig. 8 is the New Message Page and Send navigation is the most important action (b). This design will help the Elderly and Low-literacy users to avoid human error and confusion. Fig. 9 illustrates the Previous State method, where the session keeps all the state information of the application, line 2 previoussession get the previous state information and line 3 return the previous state information.

2) Button navigation: The button is the basic action mechanism to register an action. Fig. 4(e) shows the google message system button to add the message recipient's name. The “add message” recipients button symbol is not clear for the Elderly and Low-literate users moreover, there is no text information in the button. So, users might get confused about how to add new recipient's names. Fig. 8(c) shows the proposed add message recipient button, where the background color is white and the symbol is in blue color, the button has white space around it so the users will have a clear vision of the button. The proposed method will reduce human error and confusion.



Mr. Baten Gomez

Please pick up the Phone. I am trying to get you last few hours. Please call me ASAP

Not Read 10 Min Ago

How is your Son? Please Come for Dinner to Our Place. We missed you very much.

Read 6 Hours Ago

I have transfer you the Money your Account. Please che r Account and Update m

Fig. 7. Floating Navigation.

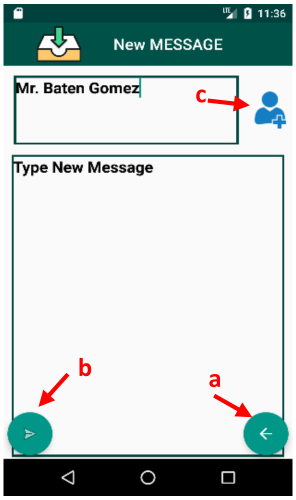


Fig. 8. Button Navigation.

```
Input:
        session is the application previous all states;
Output:
        previoussession is the previous state;

PreviousState(session)
        Do
          Go previoussession = session of previous state;
        Return previoussession;
```

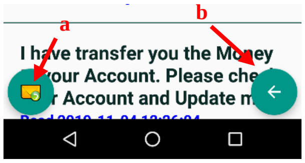
Fig. 9. Previous State Method.

E. Miss Click Avoidance

Miss click is one of the common mistakes and most of the time made by the Elderly and Low-literate users. Clicking on the interface or button requires a long press to register an action [7]. But most of the time users failed to register the action because of the proper way of pressing to the action point or shake the fingers at pressing time.

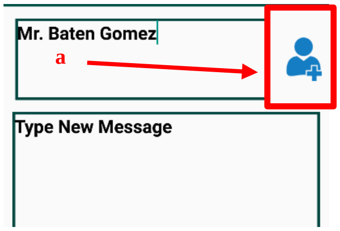
1) Action button register with additional space: To avoid the missing click of the target button, proposed additional space button design which will help the Elderly and Low-literate users to perform task smoothly. Fig. 10(a,b) shows the navigation button, both the navigation button is rounded by the red circle. Basically, the button press action gets register only when the user press exactly inside the button, but for the Elderly and Low-literate users, proposed a button which will work 5px radius outside of the button to register an action. Fig. 10(a,b) both buttons radius increase and action points work till the red circle radius, these 5px radius is invisible boundaries to the users but at action time it will work to register an action. Fig. 11(a) shows the add recipients button, the button size is small which might create a miss click of the button, so add 5px each of the button so the action button gets register very easily Fig. 11(a) red square shape.

2) Multiple and finger shake press: Most of the time Elderly and Low-literate users do not prefer to use a smartphone because they face problems interacting with the apps and the main reason they comply is multi-touch issues. As a result, they failed to register the action of the button. So, to overcome this issue proposed a method, where the user might do 2multi-tap to the action button or tap and shake finger while taped the button, the system will register the action as a single tap. Fig. 12 shows the Action Register Method, where line 1 and 2 check the button action register single tap or multiple tap or finger shake taped then the system will register the action as single tap, line 3 check the action button input boundary, if the action button is normal button, then line 4 add extra 5px (five pixels) all side of the button, as a result, the button action register capacity increase and user can register the actions very easily.



I have transfer you the Money
your Account. Please che
r Account and Update m

Fig. 10. Button Touch Action Points.



Mr. Baten Gomez

Type New Message

Fig. 11. Object Action Points.

```
Input:
        clickaction is the register all type of actions;
Output:
        registeraction is the register click action;
Variables:
        screensize is the mobile screen size;

ActionRegister(clickaction)
1.      If clickaction is register for single tap or multiple tap or shake
        and tap Then
2.        registeraction status to single tap;
3.      If buttonboundary is normal Then
4.        buttonboundary = add extra 5px boundary to all side;
5.      End
```

Fig. 12. Action Register Method.

V. Results and Discussion

Smart Phone User Interface designing is not easy for the Elderly and Low-literate users. There are many types of Smart Phone Platform is in the market like Android, iOS, Tizen, Sailfish OS, Windows, BlackBerry, Firefox OS, etc. [43]. This research has done on the Android Platform. There are many Android Messaging Applications in the Android Play Store but only a few applications are for the Elderly and Low-literate users. Table I illustrates four types of message apps first one is Google Messages [42], Raku-Raku [25], Big SMS [44] and Large Launcher [45]. Among these apps, only Google Messages [42] is open source app and all four applications are platform-dependent, which means can run only in Android and these apps are specially designed for the Elderly, moreover none of these apps are designed for the Low-literate users.

Every Smart Phone has a Message System. Table II shows the functionality of the Elderly which enhances the user experience. Multiple Tap is a critical issue to register an action. Only Raku-Raku [25] and Proposed System do not support Multiple Tap. Only Google Messages [42] is not designed for the Elderly and other apps are designed for the Elderly. Google Messages [42] and Proposed System has Navigation Consistency. The other three apps do not follow navigation consistency. Clear and Big Icons are used only in Big SMS [44] and Proposed System, other systems they use different size and color icons. Elderly and Low-literate users most of the time face problems to identify the action buttons. The Big SMS [44] system does not highlight on the button objects. The Back Button is not consistent in all the systems. Some Systems show back navigation in one page and other pages they do not have the navigations. The only proposed system continuously follows the Back Button Navigation on every page. Raku-Raku [25] and Proposed system use an algorithm to register an action to avoid miss clicks, other systems use a basic algorithm to register an action.

The time complexity depends on the flow of the algorithm. The algorithm complexity gets higher if it uses any nested operation [46]. Table III shows the comparison of the Proposed and Google Messages System [42], where *O* denotes the growth of a function and *n* is the number of steps. The Register Button Action Method, Display Message Method and Return to Previous State Method growth and time complexity are almost the same between two systems, but there are differences in the execution time because of the dependency, which related to the interaction methods and object response time.

TABLE. I. COMPARISON OF MESSAGE SYSTEM CRITERIA

	Google Messages [42]	Raku-Raku [25]	Big SMS [44]	Large Launcher [45]
Open Source	Yes	No	No	No
App code Required	Yes	Yes	Yes	Yes
Platform Dependent	Yes	Yes	Yes	Yes
Support Elderly Users	Yes	Yes	Yes	Yes
Support Low Literate Users	No	No	No	No

TABLE. II. COMPARISON OF MESSAGE SYSTEM USER INTERFACE AND FUNCTIONALITY

Criteria	Google Messages [42]	Raku-Raku [25]	Big SMS [44]	Large Launcher [45]	Proposed System
Support Multi Tap	Yes	No	Yes	Yes	No
Design UI for Elderly	No	Yes	Yes	Yes	Yes
Navigation Consistency	Yes	No	No	No	Yes
Clear Navigation Icon	No	No	Yes	No	Yes
Button Objects easy to identify	Yes	Yes	No	Yes	Yes
Previous State Navigation consistency	No	No	No	No	Yes
Algorithm to Register Action to Avoid Miss Clicks	No	Yes	No	No	Yes

TABLE. III. COMPLEXITY OF THE ALGORITHMS

	Time Complexity of Google Messages [42]	Time Complexity of Proposed System
Register Button Action Method	*O(n)*	*O(n)*
Display Message Method	*O(n)*	*O(n)*
Return to Previous State Method	*O(n)*	*O(n)*

VI. CONCLUSION

Respected Elderly and Low-literate people in society become a concern for government officials all over the world. Everyone is focusing on quality aged care and smooth communication among them to minimize the gap between the Elderly and the Present era. For better communication and interaction now days Smart Phone is the best device to work on. But all the smartphone is designed with the present trends. So, the Smart Phone trends work on the youth but when it goes to the Elderly and Low-literate users then they can see phobia or biasness is working on their mind. These groups of people scare to use the Smart Phone because of the interface and interaction methods.

Based on the research work, this paper developed a Smart Phone Messaging System only for the Elderly and Low-literate User's needs. This system focuses on the User Interface, Human and Smart Phone Interaction to enhance the usability and better Elderly and Low-literate User Experience. The proposed system is compared with other systems. The proposed system will fulfill the Elderly and Low-literate User needs, which will minimize the phobia and biasness.

A large scale of work has been carried out for the Elderly and Low-literate users. Still, the usability testing and evaluation must take place for the next step. After usability evaluation maybe more functionality or features need to be added or removed. So, the next step will do usability evaluation by the Elderly and Low-literate user and after getting the results, update the system, release it to the general public.

REFERENCES

[1] H. M. Salman , W. F. W. Ahmad, and S. Sulaiman, "Usability Evaluation of the Smartphone User Interface in Supporting Elderly Users From Experts' Perspective", IEEE Access, vol. 6, pp. 22578 - 22591, 2018.

[2] R. A. A. B. R. A.. Razak, and R. Cagadas, "Usage, Trend, Attitude, Likes and Dislikes of Elderly on New Technology Smartphone", Qualitative and Quantitative Research Review, vol. 3, no. 1, pp. 183 - 211, 2018.

[3] F. Özsungur, and O. Hazer, "Analysis of the Acceptance of Communication Technologies by Technology Acceptance Model of the Elderly: Example of Adana Province", International Journal of Eurasia Social Sciences, vol. 9, no. 31, pp. 238 - 275, 2018.

[4] C.A.L. Valk, Y. Lu, Mirana Randriambelonoro, and Jari Jessen, "Designing for technology acceptance of wearable and mobile technologies for senior citizen users", Academic Design Management Conference Proceedings (ADMC 2018), pp. 1361-1373, 2018.

[5] S. Sajjad, D. M. Khan, N. Saher, and F. Shahzad, "The Usability Analysis of Mobile Interfaces", Science International(Lahore),vol. 28, no. 2, 2016.

[6] L. Punchoojit, and N. Hongwarittorrn, "Usability Studies on Mobile User Interface Design Patterns: A Systematic Literature Review", Advances in Human-Computer Interaction, vol. 2017, pp. 1 - 22, 2017.

[7] C. Kyfonidis, and K. Renaud, "A Mobile Interface for the Older User", Proceedings of BCS Health Informatics Scotland 2016 Conference, pp. 1-9, 2016.

[8] C. D. de Oliveira, M. L. Fioravanti, R. P. de M. Fortes, and E. F. Barbosa, "Accessibility in mobile applications for elderly users: a systematic mapping", 48th Annual Frontiers in Education Conference (FIE 2018), pp. 172-180, 2018.

[9] C. Goumopoulos, I. Papa, and A. Stavrianos, "Development and Evaluation of a Mobile Application Suite for Enhancing the Social Inclusion and Well-Being of Seniors", vol. 4, no. 3, pp. 1-27, 2017.

[10] J. R. T. Araújo, R. R. T. Lima, I. M. Ferreira-Bendassolli, and K. C. de Lima, "Functional, nutritional and social factors associated with mobility limitations in the elderly: a systematic review", Salud Pública de México, vol. 5, no. 5, pp. 579-585, 2018.

[11] M. S. Al-Razgan, H. S. Al-Khalifa, and M. D. Al-Shahrani, "Heuristics for Evaluating the Usability of Mobile Launchers for Elderly People", International Conference of Design, User Experience, and Usability, pp. 415-424, 2014.

[12] B. Loureiro, and R. Rodrigues, "Design Guidelines and Design Recommendations of Multi-Touch Interfaces for Elders", The Seventh International Conference on Advances in Computer-Human Interactions, pp. 41- 47, 2014.

[13] S. K. Goel , N. Haryani , P. Tiwari , A. Jain , and P. Kuvalekar, "Smart Phone for Elderly Populace", International Journal of Research in Engineering and Technology, vol. 2, no. 10, pp. 33-35, 2013.

[14] T. van Dyk, H. Gelderblom, K. Renaud, J. van Biljon, "Mobile Phones for the Elderly: a design framework", International Development Informatics Association Conference, pp. 85-102, 2013.

[15] M. Pino, C. Granata, G. Legouverneur, M. Boulay, and A-S. Rigaud, "Assessing design features of a graphical user interface for a social assistive robot for older adults with cognitive impairment", Gerontechnology, vol. 11, no. 2, 2012.

[16] A. de L. Salgado, L. A. do AmaralRenata, P. de M. Fortes, M. H. N. Chagas, and G. Joyce, "Addressing Mobile Usability and Elderly Users: Validating Contextualized Heuristics", International Conference of Design, User Experience, and Usability, pp. 379-394, 2017.

[17] A. Hafez, Y. (Kathy) Wang, and J. Arfaa, "An Accessibility Evaluation of Social Media through Mobile Device for Elderly", Proceedings Advances in Intelligent Systems and Computing, vol. 607, pp. 179-188, 2017.

[18] D. Williams, M. A. U. Alam, S. I. Ahamed, and W.. Chu, "Considerations in Designing Human-Computer Interfaces for Elderly People", International Conference on Quality Software, pp. 372-377, 2013.

[19] H. B. Duh ,E. Y. Do, M. Billinghurst, F. Quek, V. H. Chen, "Senior-friendly technologies: interaction design for senior users", Proceeding CHI '10 Extended Abstracts on Human Factors in Computing Systems, pp. 4513-4516, 2010.

[20] R. Leunga, J. McGrenere, and P. Graf, "Age-related differences in the initial usability of mobile device icons", Journal Behaviour & Information Technology, vol. 30, no. 5, pp. 629-642, 2011.

[21] W. Qian, and W. WenDao, "Interface Design of Handheld Mobile Devices for the Older Users", International Conference on e-Education, e-Business, e-Management and e-Learning, vol. 27, pp. 185-188. 2012.

[22] I. Medhi, S. Patnaik, E. Brunskill, S.N. N. Gautama, W. Thies, and K. Toyama, "Designing mobile interfaces for novice and low-literacy users", ACM Transactions on Computer-Human Interaction, vol. 18, no. 2, 2011.

[23] J. Manuel, Dí. Bossini, and L. Moreno, "Accessibility to mobile interfaces for older people", International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, pp. 57-66, 2013.

[24] M. Pattison, and A. Stedmon, "Inclusive design and human factors: designing mobile phones for older users", PsychNology Journal, vol. 4, no. 3, pp. 267-284, 2006.

[25] K. Furuki, and Y. Kikuchi, "Approach to Commer ialization of Raku-Raku Smart Phone", Fujitsu Sci. Tech, vol. 49, no. 2, 2013.

[26] K. Renaud, R. Blignaut, and I. Venter, "Designing Mobile Phone Interfaces for Age Diversity in South Africa: “One-World” versus Diverse “Islands”", IFIP Conference on Human-Computer Interaction, pp. 1-17, 2013.

[27] N. Charness, M. Dunlop, C. Munteanu, E. Nicol, A. Oulasvirta, X. Ren, S. Sarcar, and C. Silpasuwanchai, "Rethinking Mobile Interfaces for Older Adults", Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, pp. 1131-1134, 2016.

[28] J. Gao, and A. Koronios, "Mobile Application Development for Senior Citizens", Pacific Asia Conference on Information Systems, pp. 214-225, 2010.

[29] E. Lindh, and A. John, "Designing IT for Older People", In: Exploiting the Knowledge Economy: Issues, Applications and Case Studies, PTS 1 AND 2, IOS Press, pp. 1523-1530, 2006.

[30] B. Aguiar, and R. Macário, "The need for an Elderly centred mobility policy", Transportation Research Procedia, vol. 25, pp. 4355-4369, 2017.

[31] B. W. Kiat, and W.. Chen, "Mobile Instant Messaging for the Elderly", 6th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Infoexclusion, pp. 28-37, 2015.

[32] K. Chen, Alan H. S. Chan, and S. N. H. Tsang, "Usage of Mobile Phones amongst Elderly People in Hong Kong", Proceedings of the International MultiConference of Engineers and Computer Scientists, vol. 2, 2013.

[33] K. Kalimullah, and D. Sushmitha, "Influence of Design Elements in Mobile Applications on User Experience of Elderly People", International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2017), pp. 352-359. 2017.

[34] A. L. Culén, and T. Bratteteig, "Touch-Screens and Elderly users: A Perfect Match?", The Sixth International Conference on Advances in Computer-Human Interactions, pp. 460-465, 2013.

[35] M. Faisal, M. Yusof, N. Romli, M. F. M. Yusof, "Design for Elderly Friendly: Mobile Phone Application and Design that Suitable for Elderly", International Journal of Computer Applications, vol. 95, no. 3, 2014.

[36] A. Britton, R. Setchi, and A. Marsh, "Intuitive interaction with multifunctional mobile interfaces", Journal of King Saud University – Computer and Information Sciences, vol. 25, pp. 187-196, 2013.

[37] A. Holzinger, G. Searle, and A. Nischelwitzer, "On Some Aspects of Improving Mobile Applications for the Elderly", International Conference on Universal Access in Human-Computer Interaction, pp. 923-932, 2007.

[38] S Sharma, P. Singh, R. Sharma, and A.. Mahajan, "Age based user Interface in Mobile Operating System", International Journal of Computer Science, Engineering and Applications, vol. 2, no. 2, 2012.

[39] L. G. Motti, N. Vigouroux, and P.. Gorce, "Interaction techniques for older adults using touchscreen devices: a literature review", Proceedings of the 25th Conference on l'Interaction Homme-Machine, pp. 125-135, 2013.

[40] K. Toy, E. O’Meara, R.. Kuber and S. Saulynas, "An Investigation of Ways to Support Older Adults when Using Mobile Interfaces", In iConference 2017 Proceedings, pp. 1–12, 2017.

[41] M. Kobayashi, A. Hiyama, T. Miura, C. Asakawa, M. Hirose, and T. Ifukube, "Elderly User Evaluation of Mobile Touchscreen Interactions", Conference on Human-Computer Interaction, pp. 83-99, 2011.

[42] Google Messages, November 8, 2019, https://play.google.com/store/apps/details?id=com.google.android.apps.messaging&hl=en.

[43] Top 15 Mobile Phones Operating System 2019, January 1, 2019, https://www.digitalseoguide.com/technology/top-mobile-phones-operating-systems-os.

[44] BIG SMS for Seniors, October 4, 2019, https://play.google.com /store/apps/details?id=name.kunes.android.launcher.bigmessages.

[45] Large Launcher, November 11, 2016, https://apkpure.com/large-launcher-big-font-icon/com.tbeasy.largelauncher.

[46] O. S. Pietro, H. Jun and Y. Xin, "Time complexity of evolutionary algorithms for combinatorial optimization: A decade of results", International Journal of Automation and Computing, vol. 4, pp. 218-293, 2007.

Improved Control Strategies of Electric Vehicles Charging Station based on Grid Tied PV/Battery System

Abdelilah Hassoune[1], Mohamed Khafallah[2]
Abdelouahed Mesbahi[3]
Laboratory of Energy and Electrical Systems
Hassan II University, ENSEM

Tarik Bouragba[4]
EIGSI Casablanca
Casablanca, Morocco

***Abstract*—In this paper, improved control strategies of a smart topology of EVs charging station (CS) based on grid tied PV/Battery system are designed and analyzed. The proposed strategies consist of three operating modes i.e. Pv2B; charging a battery storage buffer (BSB) of the CS from solar energy, V2G; discharging an EV battery via grid, and Pv2G; injecting the produced power from PV system into the energy distribution system. However, the BSB is connected to the PV system through a single ended primary inductor converter, the V2G operating mode is emulated by an EV lithium-ion battery tied to the grid via a high frequency full bridge inverter and a bidirectional dc/dc converter. The aim of this work is to improve the energy efficiency of the CS by using a hybrid energy system. Simulation studies are performed in Matlab/Simulink in order to operate the proposed solar CS with multiple control strategies of each case scenario based on a CS management algorithm (CSMA). To provide credible findings of this research, a low power prototype is developed in order to validate the proposed CSMA and its associated controls.**

Keywords*—Component; Electric vehicle charging station; solar energy; battery storage buffer; electrical grid; charging station management algorithm

I. Introduction

The environmental challenges and the predicted shortage of fossil fuels have made the electric vehicles (EVs) more involved in mobility sector. Yet, their development at a wider scale faces enormous constraints, especially with the charging infrastructure by which the charging station (CS) customers would require high charging power rate within a short duration time [1,2]. However, the trend toward clean and renewable energy sources (RESs), opens up a new field where a hybrid energy source will support the quick changes in power demand with less reliance on the fossil fuels. Several publications have appeared in recent years demonstrating and documenting the feasibility of novel approaches, which are based on a multi-source system gathered RESs with storage buffers [3-6]. In order to store the renewable energy and/or to overcome the power peak situation when superfast charging scenarios are set, numerous considerations must be taken into account as reported in [7]. In practical terms, the electricity distribution systems have multiple energy constraints i.e. a peak power events caused by both, a large number of plugged in EVs and connected residential or commercial alternating current (AC) loads. To frame this restriction, a distribution transformer (DT) would represent the grid, it is tied to the CS in which instantaneous power flow data would be sensed and analysed by the main controller. Recent researchers have proposed optimization methods to treat all the potential deficiencies, for instance, applying a smart scheduling of power on the supply system control to support all the charging operations based on priority levels [8-11].

The solar energy applications have been deployed very fast and drawn much attention as the case of the electric vehicle charging station (EVCS). Yet and in terms of providing a reliable energy source, the photovoltaic system is still occasionally inefficient because of the intermittent nature of solar irradiance. Nonetheless, it is possible to further enhance the PV output power by adopting new effective MPPT algorithms as discussed in [12-14]. The charging station management system (CSMS) proposed in this paper is based on a decision algorithm uses the power sensing stage in order to set the optimal operating mode for the EVCS. In terms of rapidity and stability, the CS performance would be improved using an optimization algorithm of three operating modes. The battery storage buffer (BSB) is representing the CS storage system, and it is equipped with a battery management system (BMS) which lets the CSMS to get instantaneous data i.e., the state of charge (SC) and the voltage/current. This work explores the electrical behaviour of the lithium ion battery via its charging/discharging cycles, based on that, reliable control modes are presented and analysed in detail. Thus, an experimental evaluation has given the study more credibility to draw final conclusions.

This paper is divided as follows. First, a description of the proposed topology and the operating modes are defined in Section II. Next, Section III illustrates simulation studies of the EVCS based grid tied PV/battery system. Section IV analyses experimental results of the architecture. Finally, a conclusion is presented in Section V.

II. Description of the Proposed Platform

Fig. 1 illustrates in detail the solar charging station (SCS) topology, it consists of a grid tied PV array with BSB. The solar power is injected into the station via two paths of energy flow i.e. a charging process of the BSB through a single ended primary inductor converter (SEPIC) based either MPPT algorithm to extract the maximum power from solar irradiance or voltage control mode (VCM) to avoid the overload of the connected battery. The second path is transferring electricity into the grid via dc/dc boost converter when there is a surplus of power at the PV stage or when the BSB is fully charged. However, two DC links are established to get the SCS more efficient in which numerous modes of control are implemented i.e. MPPT algorithm, VCM and current control mode (CCM).

In order to expand the research field and to obtain a quite level of profitability for both, the SCS owner and customer, a bidirectional power flow between the EV and the station is considered in the used management approach. The electrical grid is involved in the SCS via a dc/ac inverter based voltage source converter (VSC) control, the aim is to stabilize the DC link voltage via either, the PV system or the connected vehicle battery operated in V2G mode. However, Fig. 3 shows the flowchart of the main algorithm of the CSMS.

The CSMS controls the commutators of each power sources in order to apply the optimal operating mode. Then, the system generates calculated electric pulses for each converter involved in the current scenario. As it can be seen from the proposed algorithm, the presence of at least one EV at the CS will decide which operating mode will be activated either V2G, or Pv2G and Pv2B. In fact, the rush hours event is a situation when the CS would benefit as much as possible from the high selling price of the injected grid power. The three operating modes are described as follows:

A. Mode-1: Pv2G

In this mode, there is no EVs plugged in the solar CS. So, at rush hours when the selling price of kWh from CS to grid is expensive, the PV system will be given high priority to inject its power into the grid via a dc/dc boost converter and a dc/ac inverter, and as results, two power switches will be activated by the CSMS i.e. *BoostOn* and *GridOn*. The full charge of BSB will also activate this mode in order to avoid the overloading constraint (SCB_{SB} = $SC_{BSB\text{-}M}$). In the meanwhile, the control strategies provided by the CSMS will set the MPPT algorithm and the voltage source converter in order to drive the dc/dc boost converter and the dc/ac inverter, respectively.

B. Mode-2: Pv2B

In this case scenario, two power switches are activated i.e., *SepicOn* and *BSBOn*. The optimization approach of charging the CS battery explores new features in order to benefit from a maximum range of MPPT mode duration, and quick switching into the voltage regulation mode and reciprocally. The three different phases of charging the BSB are illustrated in Fig. 2.

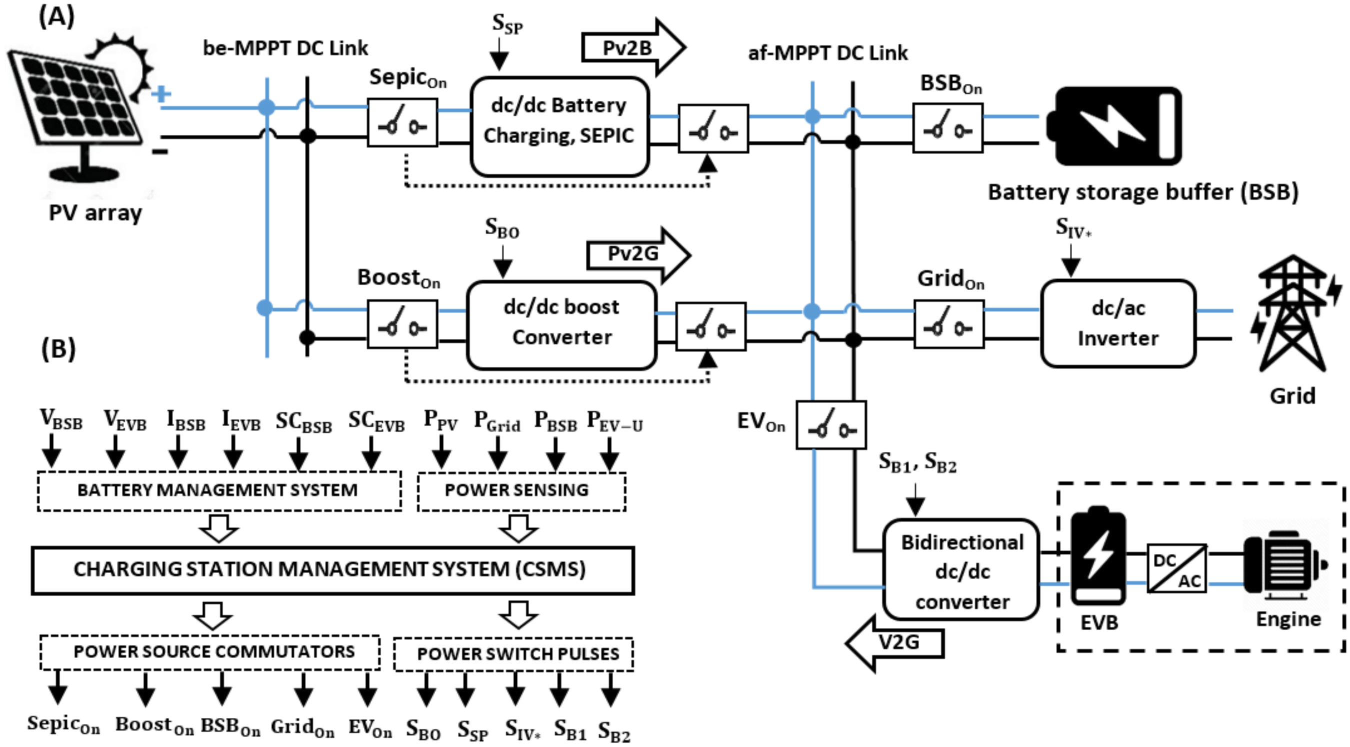

(A)
be-MPPT DC Link
S_{SP}
Pv2B
af-MPPT DC Link
+
−
$Sepic_{On}$
dc/dc Battery Charging, SEPIC
BSB_{On}
PV array
Battery storage buffer (BSB)
S_{BO}
Pv2G
S_{IV*}
$Boost_{On}$
dc/dc boost Converter
$Grid_{On}$
dc/ac Inverter
Grid
(B)
EV_{On}
S_{B1}, S_{B2}
V_{BSB} V_{EVB} I_{BSB} I_{EVB} SC_{BSB} SC_{EVB} P_{PV} P_{Grid} P_{BSB} P_{EV-U}
BATTERY MANAGEMENT SYSTEM
POWER SENSING
CHARGING STATION MANAGEMENT SYSTEM (CSMS)
Bidirectional dc/dc converter
DC
AC
EVB
Engine
V2G
POWER SOURCE COMMUTATORS
POWER SWITCH PULSES
$Sepic_{On}$ $Boost_{On}$ BSB_{On} $Grid_{On}$ EV_{On} S_{BO} S_{SP} S_{IV*} S_{B1} S_{B2}

Fig 1. The Proposed Topology (A) Solar Charging Station; (B) Power Management System.

The BSB first phase called bulk stage, is controlled by the incremental conductance (INC) MPPT algorithm by dint of the safety margin of the battery voltage which is between the allowed minimum value ($V_{BSB-min}$) and the overload maximum value ($V_{BSB-max}$). In the meanwhile, the SEPIC output current (I_{SP}) is limited to the maximum deep charge current $I_{BSB-max}$. Once V_{SP} achieved $V_{BSB-max}$, means the battery is now operating in the overcharge phase where the CSMS is switching off the MPPT algorithm and is adjusting the voltage control to the $V_{BSB-max}$. The charging current is kept increasing until it falls under $I_{BSB-min}$ where the float charge phase is reached, and in response to this change, the CSMS will readjust its reference voltage to a reduced value ($V_{BSB-float}$). This adjustment will avoid the deep self-discharge of the battery by generating a very small charging current [15].

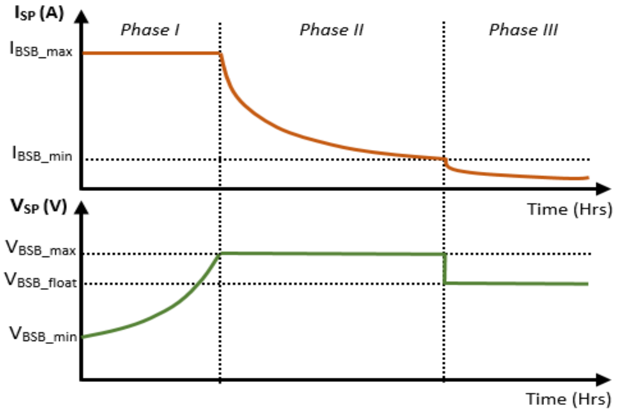


Fig 2. BSB Charging Phases.

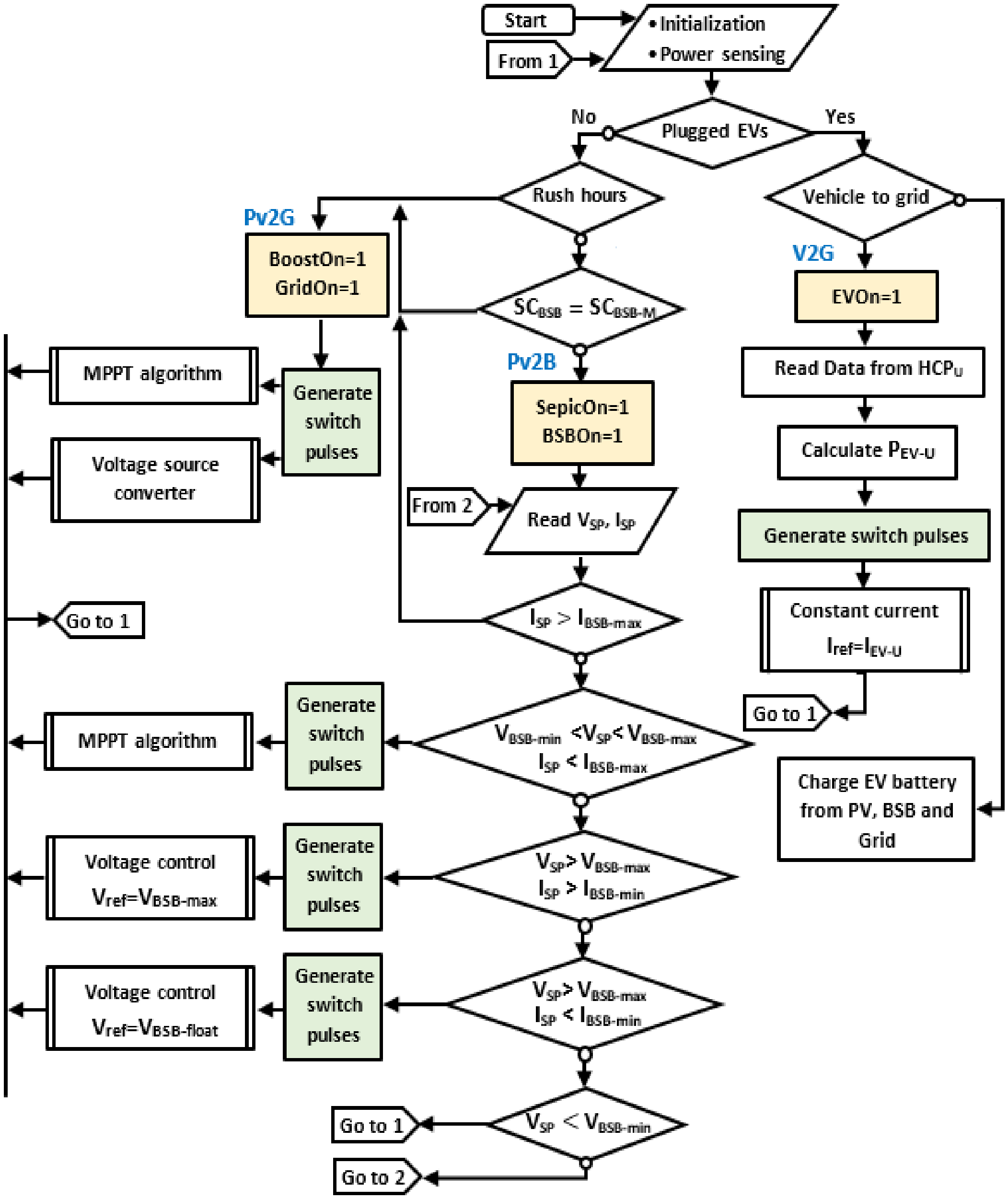


Fig 3. Flowchart of the Charging Station Management Algorithm (CSMA).

C. Mode-3: V2G

Once an EV battery is plugged into a CS outlet, the CSMS will activate the *EVOn* and will extract the data from the human control panel of each vehicle user (HCP_U).

The power to be injected into the grid from EV battery is expressed as follows:

$$P_{EV-U} = \frac{(SC_{EVB} - SC_{EVB^*}) \times BC_{EVB}}{PT_{EV}} \tag{1}$$

Where P_{EV-U} is the injected power from an EV battery, SC_{EVB} is the initial state of charge (SC) of a battery, SC_{EVB^*} is the final SC set by the vehicle user, BC_{EVB} is the battery capacity of the EV and PT_{EV} is the plugged time. The management system will calculate and update the reference current used to drive the bidirectional charger switches as expressed in the following equation:

$$I_{EV-U} = \frac{P_{EV-U}}{V_{EVB}} \tag{2}$$

Where I_{EV-U} is the reference current set by the CSMS, and V_{EVB} is the voltage of an EV battery. The control of the battery discharger consists of two kinds of strategy i.e., constant voltage and constant current [16-18]. When a vehicle battery is connected, it would be recommended at first to start the discharging process with a constant voltage control, and then to switch the control to a constant current (I_{ref}=I_{EV-U}). Fig. 4 depicts the battery discharger operating under a constant current control, a PI controller is used to regulate the margin error between the discharging current and its reference value, followed by a pulse width modulation block to generate the required pulses i.e., S_{B1} and S_{B2} [19-21].

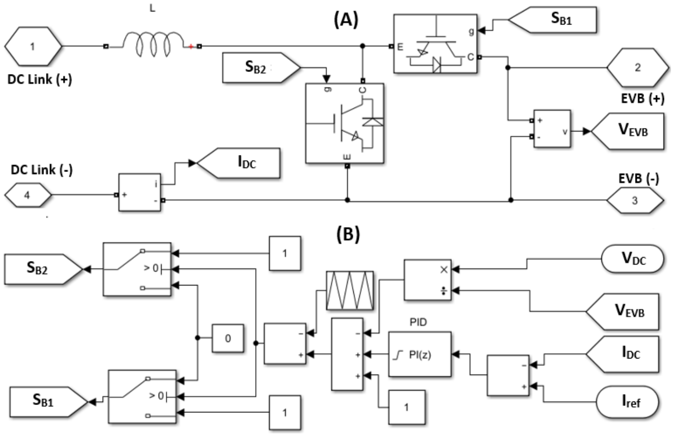


Fig 4. EV Battery Discharger (A) Bidirectional dc/dc Converter (B) Constant Current Control.

III. Simulation Results

The simulation study of the proposed topology is designed through a 7.68 kWp ISF-240 PV system prototype, in (8×4) configuration series/parallel, respectively. The af-MPPT DC link voltage is adapted at 500 V in order to ensure an eventual power injection into the energy distribution system. Thus, the grid is connected via a ac/dc inverter in order to operate the SCS in V2G mode. Fig. 6 illustrates the modelled block diagram in Matlab/Simulink.

To frame the control strategies, the system has been modelled and designed on Matlab/Simulink software, which is carried out three case scenarios:

- Case 1 (Pv2B): BSB charging mode via SEPIC controlled by MPPT algorithm/VCM.
- Case 2 (Pv2G): Power injection mode into grid from solar energy via a boost converter and a dc/ac inverter controlled by a MPPT algorithm and a VSC control, respectively.
- Case 3 (V2G): Vehicle to grid mode via a bidirectional dc/dc converter and a dc/ac inverter controlled by a constant current control and a VSC control, respectively.

The proposed block diagram is developed to emulate a part of a CS. However, the simulation is based on multiple scenarios, in which the previous cases will be tested. Fig. 5 depicts the adopted climatic scenario.

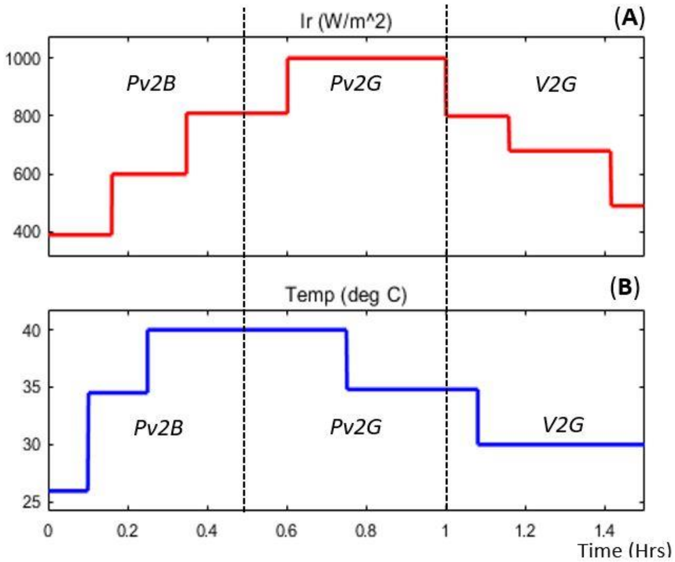


Fig 5. Climatic Scenario Versus Time (A) Irradiation; (B) Temperature.

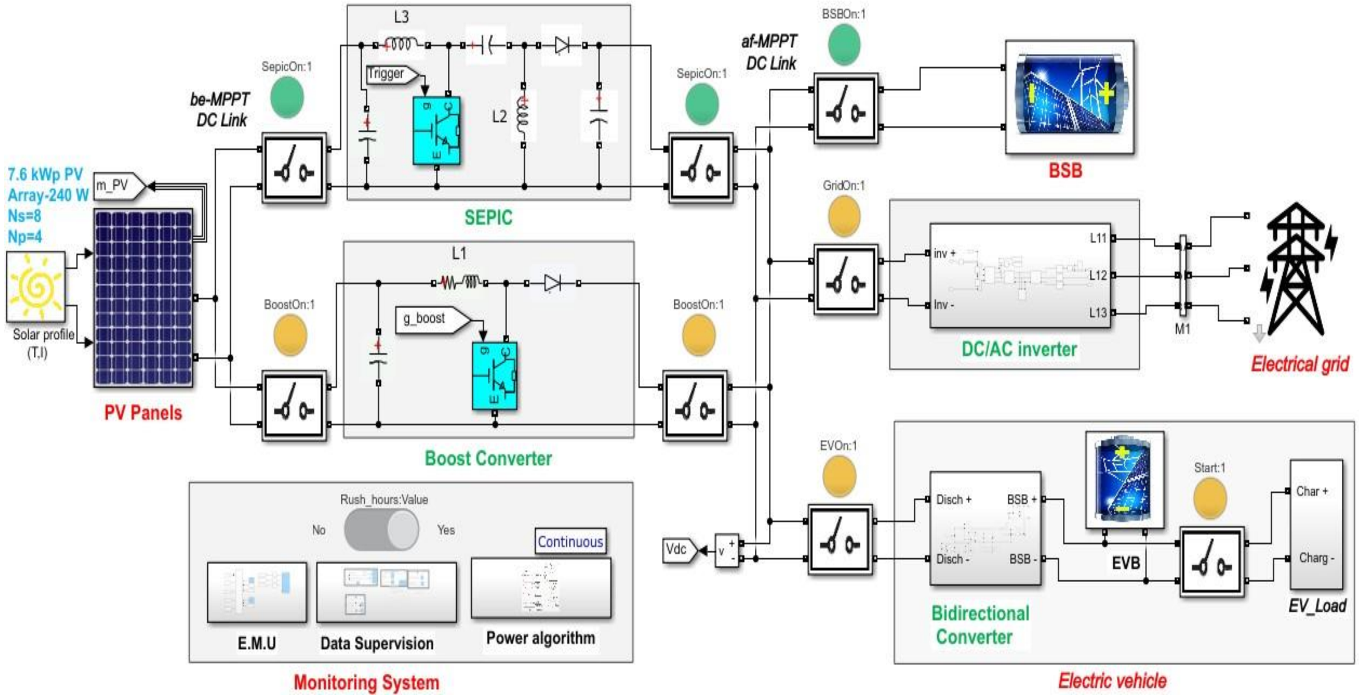


Fig 6. Proposed SCS Topology in Matlab/Simulink.

A. Case-1: PV to BSB (Pv2B)

During a sunny day with a high level of irradiation, the Pv2B mode is given high priority to gain from solar energy to load the CS battery (BSB). At this case, the battery charger is operated under one of the two modes of control, where numerous constraints are treated to ensure a reliable charging process. Fig. 7 shows the used timeline of a climatic scenario and the charging power injected into the BSB.

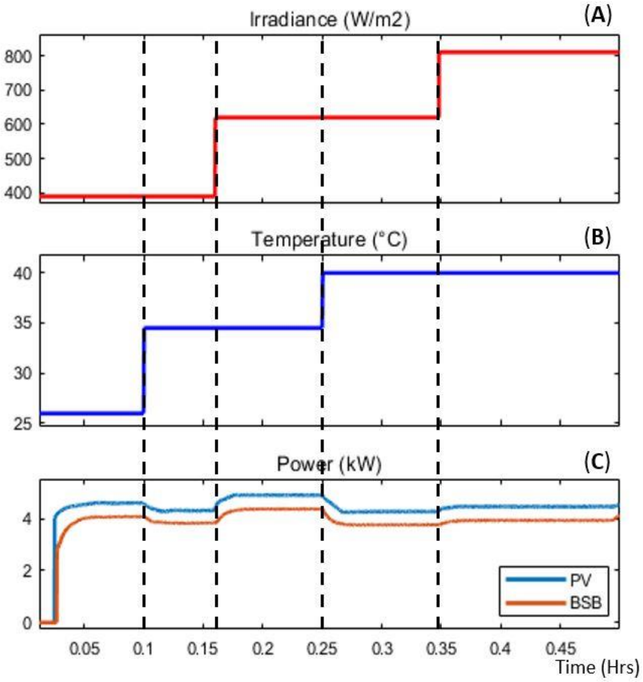


Fig 7. Curves of the Pv2B mode versus Time (A) Irradiation; (B) Temperature; (C) PV/BSB Power.

For t = 0-0.1Hrs, the system aims to extract the maximum power from 400W/m^2; 26°C. After 0.1Hrs, the temperature is up to 34.7°C and as result, the PV power is decreased to 4.1kW, thus the temperature is inversely proportional to the generated power from solar array. In these circumstances, the system achieves an energy efficiency close to 92%, in which the energy losses are mainly related to passive converter components. After t = 0.16Hrs, the irradiance is increased to 610W/m^2, so the PV power is up to 4.6kW. Therefore, the power is proportional to the changes of solar irradiance. Besides, Fig. 8 shows the charging process of the BSB via current and SC curves.

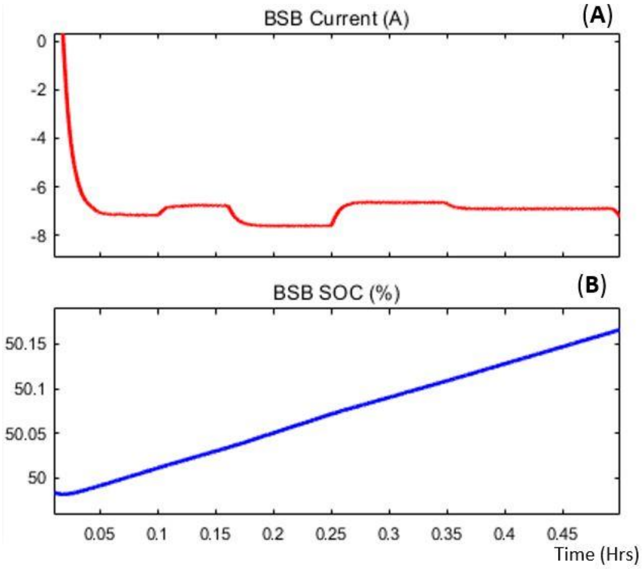


Fig 8. Curves of the BSB Charging Phase versus Time (A) Current; (B) SC.

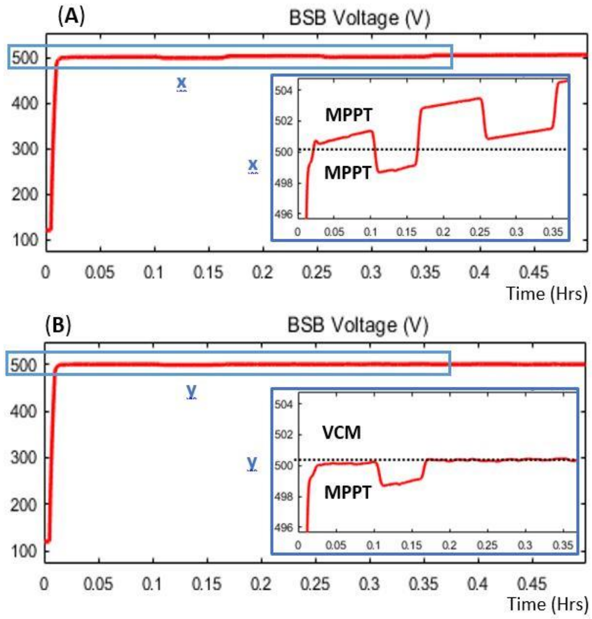


Fig 9. Curves of the BSB Charging Voltage versus Time (A) without CSMS; (B) with CSMS.

In this operating mode, the CSMS adds further complexities and constraints to the control strategy in order to achieve safety charging/discharging processes of the BSB. Therefore, the algorithm will avoid the depth of discharge state of the CS battery and to not operate it in the harmful phase of saturation. As it can be seen in the previous results, the changes of solar irradiation affect the curve slope of the BSB SC which is featured by its inclination rate, thus, the charging rate is not constant throughout the charging phase. However, Fig. 9 illustrates the results of the improved control strategy composed by voltage control mode and MPPT algorithm, each one of them is selected by the CSMS one at a time.

The Pv2B mode uses SEPIC as a battery charger of the BSB. The battery performance degradation is occurred when the charger output voltage surpasses the maximum rate of the BSB voltage, the MPPT algorithm control can cause this case scenario during the detection phase of the MPP as depicted in Fig. 9A. So, the management algorithm will fix the overloading dilemma during a charging phase of the battery as shown in Fig. 9B.

B. Case-2: PV to Grid (Pv2G)

Since electricity cannot be stored on a large scale, this mode is established to get more financial revenues for the CS owner based on RESs. Yet, the intermittent still very marked in the day/night alternation for solar power where this case is only possible in daylight. The electrical grid is equipped with an accurate mode of control to stabilize the injected power. So, the process is started with a dc/dc boost converter based MPPT algorithm, and a dc/ac inverter based VSC control. In order to maintain the DC link stable, the VSC is operated under Phase-locked loop (PLL), the loop system generates an output signal whose phase is related to the phase of a reference signal. However, Fig. 10 illustrates the injected power into the grid following the adopted climatic scenario.

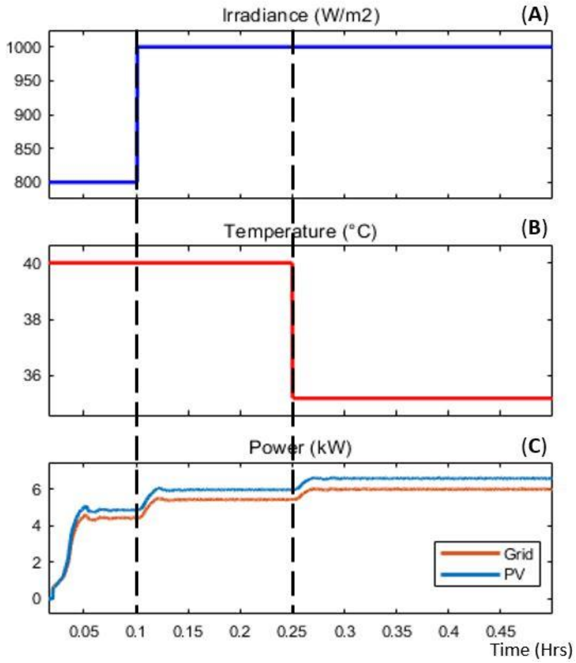


Fig 10. Curves of Pv2G Operating mode versus Time (A) Irradiation; (B) Temperature; (C) Power.

The VSC is composed by PLL followed by current regulator and PWM controller, the output is generated pulses sent to the inverter switches. During the simulation time, three different scenarios are tested on the PV system where high irradiation and low temperature made a best energy efficiency for the SCS.

C. Case-3: Vehicle battery to Grid (V2G)

The EVs can be plugged into the electrical grid via home or other bidirectional power slot. This work presents a study case of a small parking place equipped with multiple charging points, they have been provided with an HCP_U so the customer could set his requirements in terms of power demand. A proposed HCP_U data of three EVs is presented in Table I.

In order to apply the reference power on the bidirectional dc/dc converter control with a quite level of efficiency, the private BMS of each EV will provide the CSMS with the required data form each EV user through the HCP_U. Fig. 11 and 12 show the curves of power, voltage and current of the grid versus the variations of data from the user control panel. The injected power into the grid is accordingly updated to each power scenario.

TABLE I. HCP_U DATA OF THREE EVS OPERATED IN V2G OPERATING MODE

EV	Brand	Battery Capacity (kWh)	Plugged Time (Hrs)	Initial SC (%)	Final SC (%)	EV demand (kW)
1	Nissan Leaf 2	40	0.15	93	81	32
2	Kia e-Niro	64	0.2	87	79	25.6
3	Tesla Model X	75	0.15	95	85	50

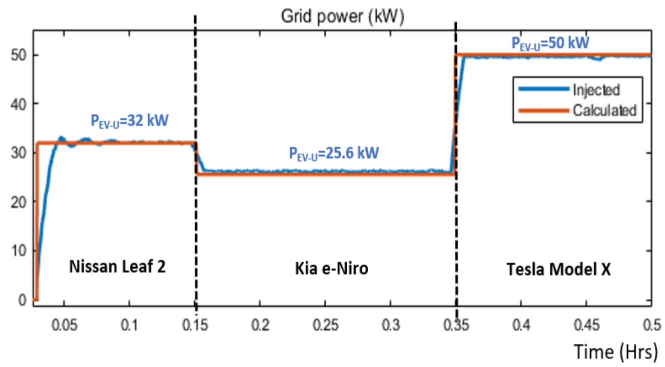


Fig 11. Curve of Grid Power versus Time.

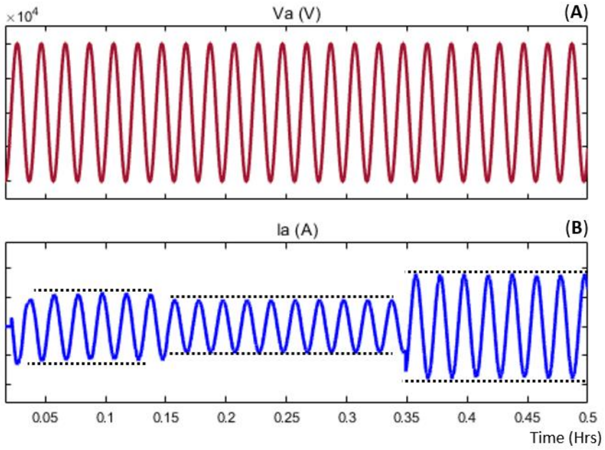


Fig 12. Grid Curves versus Time (A) Voltage; (B) Current.

In this mode, the bidirectional charger of the EV battery is operated in discharging mode, it is controlled by a current control mode based on PI block that regulates the stability, the rapidity and the accuracy of the discharging process.

IV. Experimental Evaluation

A laboratory prototype of the proposed topology is set to test the effectiveness of the control strategies. The PV output power is controlled by Texas instrument Solar Explorer Kit. It puts up a flexible low voltage platform to assess the C2000 microcontroller family for solar power applications. Fig. 13 illustrates the printed circuit board, in which the proposed control strategies will be verified.

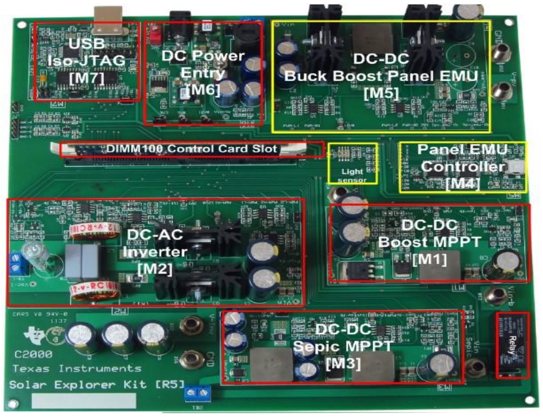


Fig 13. Macro Blocks of Solar Explorer Kit.

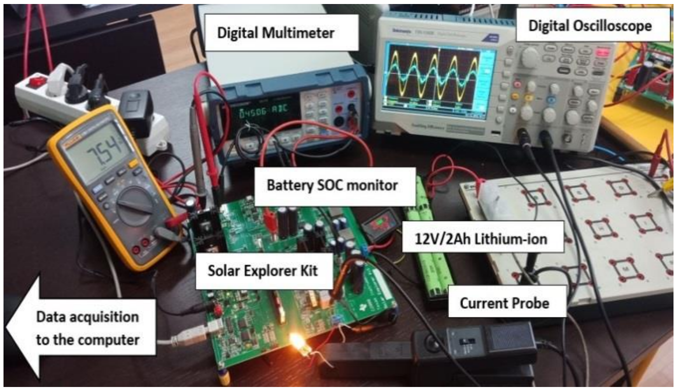


Fig 14. Experimental Prototype of the Solar Charging Station.

The circuit board consists of F28035 piccolo control card in order to drive the dc/dc boost converter, the SEPIC and the inverter. The PV array is emulated via a synchronous buck boost stage controlled by Piccolo-A F28027 card.

The prototype is set up to emulate the multiple stages of dc/dc and of dc/ac conversion along with a real-time processing to run the various kinds of control. A PV emulator is built onto the board with a buck/boost power stage using light sensor as data input. To complete the demonstration, a lithium-ion battery is used to perform the 12V/2Ah BSB. Fig. 14 shows the experimental prototype of the solar CS topology.

The PV-grid tied inverter uses a dc/dc boost converter operated at 100kHz to increase and to stabilize the output voltage at 30V (af-MPPT DC link). The DC link voltage at the dc/ac inverter input is required to be higher than the maximum AC voltage. Fig. 15 shows the curves of voltage and current at the inverter output stage operated at 20kHz. As it can be seen, both the voltage and the current are set to be operated at the grid frequency (50Hz). To highlight the control strategy of the injection mode, several modulation schemes are analyzed in order to feed current into the grid through a high frequency full bridge inverter. This control strategy is basically based on unipolar modulation used to switch alternate legs depending on which sine half of the AC signal is being generated. Fig. 16 illustrates the AC grid power. The inverter output is connected to LCL filter to improve the waveforms of grid voltage and current, thus the grid power in the channel *M* (red) obtained 16.1VA.

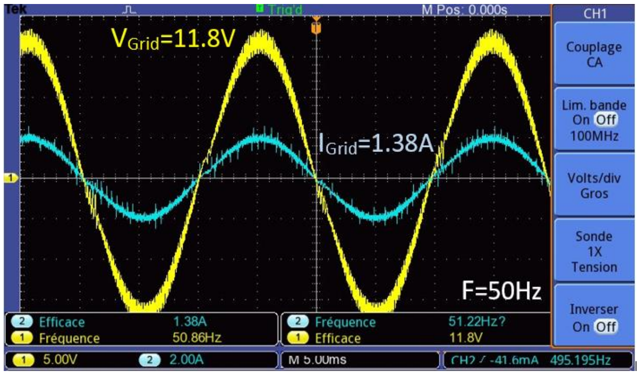


Fig 15. Experimental AC Voltage (CH1) and AC Current (CH2) at Grid side versus Time.

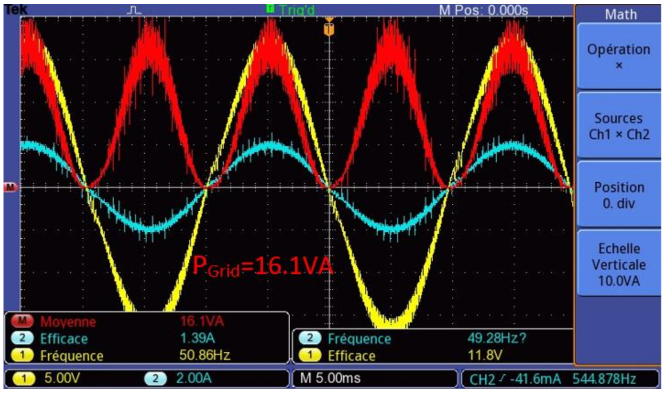


Fig 16. Experimental AC grid Power versus Time (CHM).

In this mode, a synchronous buck/boost stage is used to emulate the PV array, the input voltage is provided from a DC power block, where the kit is supplied by a 24V/2.5A. Fig. 17 and 18 show the experimental DC power, voltage and current at 0.5kW/m^2 of irradiation.

At 0.5kW/m^2, the PV emulator delivers 18.8W and the power at the inverter output stage is 16.18VA. Further experiences are done under various luminance ratios in order to create concrete climatic scenarios. Table II summarizes the experimental results of the injection mode into the grid under various rates of irradiance (0.2-1kW/m^2).

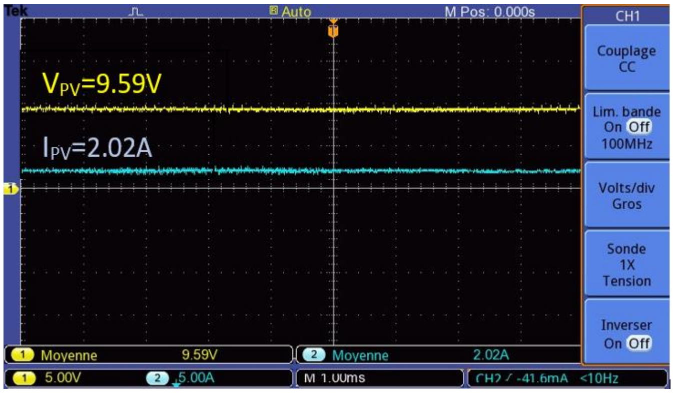


Fig 17. Experimental DC Voltage (CH1) and Current (CH2) versus Time at 0.5kW/m$^{2.}$

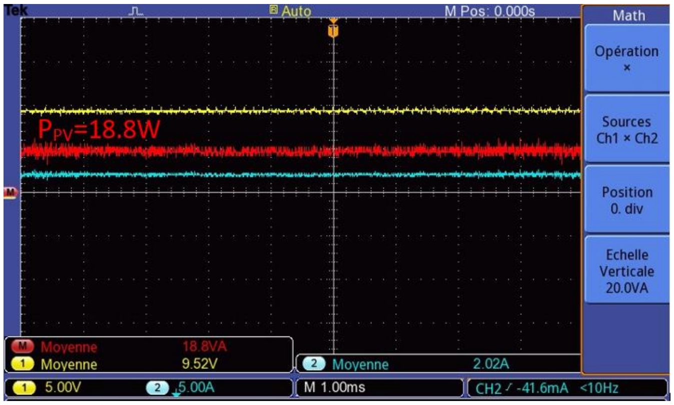


Fig 18. Experimental DC Power (Red) versus Time at 0.5kW/m^2.

TABLE II. EXPERIMENTAL RESULTS OF DIFFERENT RATIOS OF SOLAR IRRADIANCE

Luminance Ratio (kW/m^2)	Power at MPP (Watts)	Voltage at MPP (Volts)	Grid power (VA)
1	36.02	18.46	33.17
0.9	32.42	16.42	29.54
0.8	28.82	14.68	25.95
0.7	25.22	12.77	22.23
0.6	21.61	10.98	18.81
0.5	18.80	9.516	16.18
0.4	14.41	7.363	12.29
0.3	10.81	5.473	9.31
0.2	7.205	3.67	6.23

TABLE III. POWER STAGE PARAMETERS OF THE SEPIC

Voltage (V)		Current (A)		Power rating max (W)	Frequency (kHz)
Input	*Output*	*Input*	*Output*		
0-30	10 - 16	0-3.5	0-3.5	50	200

During the Pv2B operating mode, some precautions are established in order to get the system matched the required case scenario. Table III expresses the features of the allowed margin of the connected load.

A 12V/2Ah lithium-ion battery is chosen to test the 50W PV emulator. However, the battery charger is operated under a hybrid strategy of control. Fig. 19 and 20 show the voltage, the current, and the power at the input and at the output of SEPIC at 0.2kW/m^2.

As follows from the figures shown above, the SEPIC is operated at buck mode due to the low irradiance rate applied in the PV emulator input (0.2kW/m^2). From the characteristics of the adopted battery, an overcharge voltage is fixed to V_{OC}=15V, so the condition in which the CSMA would switch the control from MPPT algorithm to VCM is when the BSB voltage exceeds 0.95 V_{OC} (14.25V), such case is depicted in Fig. 20, where V_{BAT} still at 0.86 V_{OC} (12.7V). In this case, the dc/dc converter is controlled by MPPT algorithm.

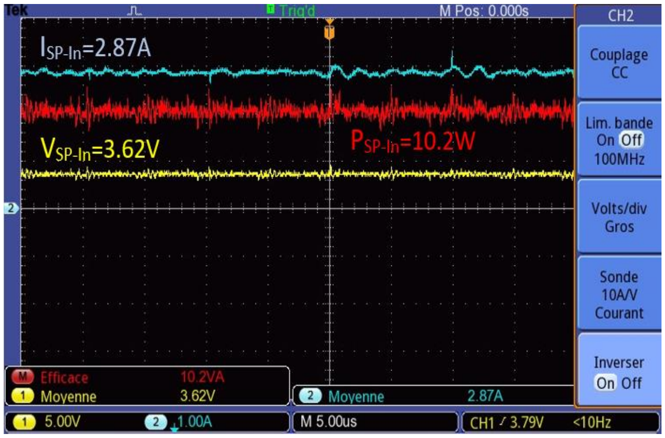


Fig 19. Experimental DC Voltage, Current and Power at SEPIC Input versus Time.

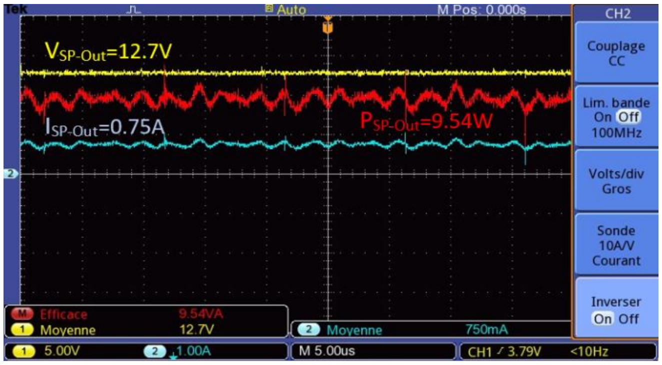


Fig 20. Experimental DC Voltage, Current and Power at SEPIC Output versus Time.

V. Conclusion

In this paper, a solar charging station for EVs is presented in detail to design and to analyze three power flow scenarios i.e., Pv2B, Pv2G and V2G. In order to maintain a high level of accuracy and stability of each operating mode, hybrid control strategies were validated by simulation and experimental tests. The reliability and the flexibility of this approach are achieved by a charging station management algorithm. From the proposed topology, the management algorithm is productive while considering some criteria, such as the capacity of the BSB, the installed power of PV array, and of grid transformer.

The accuracy and the stability of each operation mode has been validated by simulation results of 7.68 kW power system performed in Matlab/Simulink. Thus the findings are analyzed with experimental evaluations using a low power solar kit. In terms of decreasing the charging cost from over-relying on the grid and for a smooth integration of eventual other renewable energy sources, this optimization approach obtained gainful outcomes. However, it is also important to analyse the economic and the reliability aspects of the proposed system based on hybrid energy sources, which can also be treated as the future scope of this work.

References

[1] B. Dolter and N. Rivers, “The cost of decarbonizing the Canadian electricity system,” Energy Policy, vol. 113, pp. 135–148, Feb. 2018.

[2] A. R. Bhatti, Z. Salam, M. J. B. A. Aziz, K. P. Yee, and R. H. Ashique, “Electric vehicles charging using photovoltaic: Status and technological review,” Renewable and Sustainable Energy Reviews, vol. 54, pp. 34–47, Feb. 2016.

[3] H. Jakir, N. Sakib, E. Hossain and R. Bayindir, “Modelling and Simulation of Solar Plant and Storage System: A Step to Microgrid Technology,” International Journal of Renewable Energy Research, vol. 7, no. 2, pp. 723–737, 2017.

[4] J. A. Domínguez-Navarro, R. Dufo-López, J. M. Yusta-Loyo, J. S. Artal-Sevil, and J. L. Bernal-Agustín, “Design of an electric vehicle fast-charging station with integration of renewable energy and storage systems,” International Journal of Electrical Power & Energy Systems, vol. 105, pp. 46–58, Feb. 2019.

[5] J. Lee and G.-L. Park, “Dual battery management for renewable energy integration in EV charging stations,” Neurocomputing, vol. 148, pp. 181–186, Jan. 2015.

[6] I. A. Nienhueser and Y. Qiu, “Economic and environmental impacts of providing renewable energy for electric vehicle charging – A choice experiment study,” Applied Energy, vol. 180, pp. 256–268, Oct. 2016.

[7] P. Goli and W. Shireen, “PV powered smart charging station for PHEVs,” Renewable Energy, vol. 66, pp. 280–287, Jun. 2014.

[8] A. R. Bhatti, S. Zainal, J. B. A. A. Mohd and P. Y. Kong, “A Comprehensive Overview of Electric Vehicle Charging Using Renewable Energy,” International Journal of Power Electronics and Drive Systems, vol. 7, no. 1, pp. 114-123, 2016.

[9] A. Hassoune, M. Khafallah, A. Mesbahi, L. Benaaouinate, and T. Bouragba, “Control Strategies of a Smart Topology of EVs Charging Station Based Grid Tied RES-Battery,” International Review of Electrical Engineering (IREE), vol. 13, no. 5, pp. 385-396, Oct. 2018.

[10] H. Fathabadi, “Novel solar powered electric vehicle charging station with the capability of vehicle-to-grid,” Solar Energy, vol. 142, pp. 136–143, Jan. 2017.

[11] A. Luo, Q. Xu, F. Ma and Y. Chen, "Overview of power quality analysis and control technology for the smart grid," in Journal of Modern Power Systems and Clean Energy, vol. 4, no. 1, pp. 1-9, January 2016.

[12] Y. Cheddadi, F. Errahimi, and N. Es-sbai, “Design and verification of photovoltaic MPPT algorithm as an automotive-based embedded software,” Solar Energy, vol. 171, pp. 414–425, Sep. 2018.

[13] Y. Aljarhizi, A. Hassoune, and E. M. Al Ibrahmi, “Control Management System of a Lithium-ion Battery Charger Based MPPT algorithm and Voltage Control,” 2019 5th International Conference on Optimization and Applications (ICOA), Apr. 2019.

[14] R. Bradai, R. Boukenoui, A. Kheldoun, H. Salhi, M. Ghanes, J.-P. Barbot, and A. Mellit, “Experimental assessment of new fast MPPT algorithm for PV systems under non-uniform irradiance conditions,” Applied Energy, vol. 199, pp. 416–429, Aug. 2017.

[15] J. P. Torreglosa, P. García-Triviño, L. M. Fernández-Ramirez, and F. Jurado, “Decentralized energy management strategy based on predictive controllers for a medium voltage direct current photovoltaic electric vehicle charging station,” Energy Conversion and Management, vol. 108, pp. 1–13, Jan. 2016.

[16] J. Caballero, J. Chinchilla, and J. Rosero Garcia, “Performance Testing and Power Quality of DC Semi-Fast Chargers of Electric Vehicles (EVs) for Public Transportation: A Case Study,” International Review of Electrical Engineering (IREE), vol. 11, no. 6, pp. 579, Dec. 2016.

[17] A. Hassoune, M. Khafallah, A. Mesbahi, and T. Bouragba, “Power Management Strategies of Electric Vehicle Charging Station Based Grid Tied PV-Battery System,” International Journal of Renewable Energy Research (IJRER), vol. 8, no. 2, pp. 851-860, Jun. 2018.

[18] S. Chalise, J. Sternhagen, T. M. Hansen, and R. Tonkoski, “Energy management of remote microgrids considering battery lifetime,” The Electricity Journal, vol. 29, no. 6, pp. 1–10, Jul. 2016.

[19] A. Mendoza-Torres, N. Visairo, C. Nuñez, J. Armenta, E. Rodríguez, and I. Cervantes, “Switching rule for a bidirectional DC/DC converter in an electric vehicle,” Control Engineering Practice, vol. 82, pp. 108–117, Jan. 2019.

[20] W. Wojtkowski, “Digital control of a bidirectional DC / DC converter for automotive applications,” IFAC-PapersOnLine, vol. 51, no. 6, pp. 113–118, 2018.

[21] A. Hassoune, M. Khafallah, A. Mesbahi, and T. Bouragba, “An Improved Approach of Control for a Battery Charger Based Forward Converter and SEPIC,” 2018 6th International Renewable and Sustainable Energy Conference (IRSEC), Dec. 2018.

Problem based Learning with Information and Communications Technology Support: An Experience in the Teaching-Learning of Matrix Algebra

Norka Bedregal-Alpaca[1], Olha Sharhorodska[2]
Victor Corneko-Aparicio[4]
Departamento Académico de Ingeniería de Sistemas e Informática, Universidad Nacional de San Agustín de Arequipa, Arequipa, Perú

Doris Tupacyupanqui-Jaen[3]
Departamento Académico de Matemáticas
Universidad Nacional de San Agustín de Arequipa
Arequipa, Perú

***Abstract*—Students and teachers face problems in the teaching-learning processes of matrix algebra, due to the level of abstraction required, the difficulty of calculation and the way in which the contents are presented. Problem-based Learning (PBL) arises as a solution to this problem, as it contextualizes the contents in everyday life, allows students to actively build that knowledge and contributes to the development of skills. The proposal describes a didactic sequence based on PBL, which uses cooperative techniques and MATLAB, as instruments that facilitate the resolution of problems close to the student experience. The features of the Moodle platform are used to support the face-to-face educational process. The perception of students, in relation to the activity shows that 83% believe that it contributed to the understanding of the topics covered and 79% think that it allowed them to develop their creativity and capacity for expression.**

***Keywords*—Problem-based learning; cooperative techniques; constructionism; matrix algebra**

I. Introduction

In 1850, J. Sylvester first used the term "matrix" and defined it as a rectangular arrangement of terms. Since then, matrix algebra has contributed to the development of many areas of knowledge: qualitative theory of differential equations, cryptography, mathematical optimization, decision theory, robotics, astronomy, etc.

The traditional way in which algebra is taught has resulted that many students finding it difficult to learn so they experience a rejection towards it, a rejection that is transferred to mathematics in general [1], [2]. It is necessary to consider, in the teaching-learning processes, consider situations in which the student can systematically move between the concrete and the abstract; so that algebraic concepts and processes make sense. In the problem experienced by the student, the level of abstraction required, the way in which the contents are presented, the lack of handling of the conceptual prerequisites (basic operations, rows, columns, variables, function) and the lack of contextualization with situations close to reality. In [3] and [4] it is proposed to use problem solving and project development as teaching-learning strategies in linear algebra courses.

Thus, Problem Based Learning (PBL) emerges as a solution to the problem described, because it contextualizes the contents in everyday life, allows students to actively build knowledge and contributes to the development of communication and communication skills of joint work.

The experience described below was aimed at showing that the integration of PBL and the use of Information and Communication Technologies (ICT) favors the prominence of the student. By using a symbolic calculation system (MatLab) in the execution of routine tasks, the necessary space is achieved for the student to focus his efforts on more general mental processes, develop his reasoning ability and creativity, as well as another set of skills and capabilities that will serve you for your professional development.

II. Conceptual Framework

A. Problem-Based Learning (PBL)

It is a teaching-learning method characterized by promoting self-directed learning and critical thinking, part of a problem (designed or outlined by the teacher) that the student must solve in order to develop skills and competencies previously defined. The problem situations on which the ABP is based are real-world scenarios, the teacher models them or adapts them to give the student the possibility to investigate or experiment on the nature of these phenomena or daily activities.

In [5], it is argued that the starting point for the acquisition and integration of new knowledge is to use problems. There are several ways to work with PBL, [6] summarizes the steps students should take for their application: (a) reading and analyzing the problem scenario, (b) brainstorming, (c) listing what is known, (d) and a list of what is not it is known, (e) make a list of what needs to be done to solve the problem, (f) define the problem, (g) obtain information and (h) present results.

B. Basic Mathematic

Basic Mathematics is a subject taught in the first semester in most professional schools of the Peruvian university system. Among other topics, it deals with mathematical objects such as: real numbers, relationships, functions, vectors, and matrices.

When developing matrix algebraic operations, students capture the logical sequence to perform operations without further difficulty except when the calculations are large, but when asked to analyze the information, they are confused and exposed wrong opinions.

To solve the analysis of the information will work on solving problems that will be modeled by matrices, such as a graphic figurative system, that represented reality graphically and symbolically to make the problem more understandable, it is intended so that the student manages to appropriate the concept.

III. Methodology

The experience was held at the Professional School of Mechanical Engineering of the National University of San Agustín de Arequipa (Peru) during the 2017-A period. The population under analysis consists of 30 students enrolled in the “Basic Mathematics” subject.

This study is part of a field investigation with a quasi-experimental design (a control group has not been included), quantitative and phenomenological type. Students to carry out the activities are organized based on cooperative learning techniques.

A questionnaire has been used to collect the students' perception of the experience.

IV. Description of the Experience

A. Context

The subject Basic Mathematics is taught in the first semester of the career; equivalent to 4 academic credits corresponding to 4 hours of theory and 2 hours of practice. In the development of the subject, in the theoretical hours the participatory master class has been used as a form of organization and in the practical hours some cooperative work activities were implemented.

In particular, PBL-based activities were implemented for the development of the topic “Operations with arrays and applications”.

B. Features

The experience was conducted in 3 theoretical-practical sessions, communication between students and the teacher, as well as information management and survey were conducted through different functionalities of the Moodle platform (email, forum, wiki, file, task, survey, questionnaire, etc.).

The groups were formed on the basis of the principles of Cooperative Learning (CL). Using the grades obtained in the evaluation, each team was made up of a well-performing, one low-performing, and two mid-level students. With these premises, 5 teams were formed with 4 members and 2 teams of 5 members.

To organize group work, roles within the group were defined; following [6, 7] four roles were considered: coordinator, academic manager, editor-in-chief and creative manager. These roles were rotating, so that each team member was able to perform each of the defined roles.

To model the problems, the recommendations of [8], who say that an adequate problem to work from the ABP, should be based on real problems, preferably multidisciplinary, be related to the teaching objectives of the subject, to be open, topical, complex and appropriate to the cognitive and motivational level of students.

MatLab was chosen for the execution of calculations because of their availability in the computer lab and because the operations are expressed in a similar way as in mathematics.

The implementation of the virtual classroom followed the instructional design presented in [9].

C. The Experience Stages

1) Stage 1: Presentation of the theoretical basis and the characteristics of the work to be performed

Face-to-face session in which a didactic sequence was worked on to present the matrix operations (sum, difference, multiplication of a scalar by a matrix, multiplication of matrices).

In this sequence the concepts were presented, the operations were explained, exercises were solved and some problems were solved. To consolidate the work done, a link was configured in the virtual classroom through which they could access a set of 4 problems that had to be solved under the PBL methodology.

An additional 10 minutes were used to explain the cooperative characteristics of the work, the PBL and the formation of groups:

- The characteristics of a cooperative work, the functions of each of the roles within the group and the need to rotate the roles were explained.
- General information was given on the 8 steps to apply PBL.
- It was reported that some Moodle functionalities (forum, chat and wiki) that could be used to work the problems had been configured for each group.
- It was proposed that in the next face-to-face session one of the problems would be solved as an example.

All this information was placed in the virtual classroom of the Moodle platform, in addition material related to the ABP was placed. This decision was made taking into consideration of the results obtained in relation to the acceptance of the use of the platform [10] in previous work.

2) Stage 2: Presentation and solution of the problem

According to the schedule, a theoretical-practical class was targeted to the solution of one of the proposed problems.

Problem Presentation

Ana the secretary of the "Blue Earth" Research project is in charge of the purchase of the office material.

She had received three catalogs of office supplies stores, before ordering the purchase you decide to do a price study to save.

Ana has to decide on a supplier, to do so she takes a sample, the prices of five items in the three suppliers and will estimate the expense she would have following last year's orders in the same month.

For the study, Ana considers the prices of: notebooks, pen boxes, spare parts of ink cartridges, staple boxes and thousands of Bond paper.

With the first supplier, the prices are S / 15.00 the notebooks, S / 30.00 Box of 50 Faber Castell Pens, S / 2.70 box of Staples 26/6 x 5000 Arty, S / 27.90 c / u HP 664 Tri-Color Ink SKU: 250893-1, S / 9.60 75g Bond A4 paper. 500 sheets

The second supplier has the prices S / 15.50 the notebooks, S / 30.50 Box of 50 Faber Castell Pens, S / 3.10 box of Staples 26/6 x 5000 Arty, S / 27.00 c / u Ink HP 664 Tricolor SKU: 250893-1, S / 9.80 75g Bond A4 paper. 500 sheets

The third supplier has the prices S / 13.80 the notebooks, S / 29.50 Box of 50 Faber Castell Pens, S / 3.20 box of Staples 26/6 x 5000 Arty, S / 27.60 c / u Ink HP 664 Tricolor SKU: 250893-1, S / 9.30 75g Bond A4 paper. 500 sheets.

Orders placed the previous year in the months of October, November and December are presented in Table I.

At this point the question was asked: What decision should Anne make?

Problem solution:

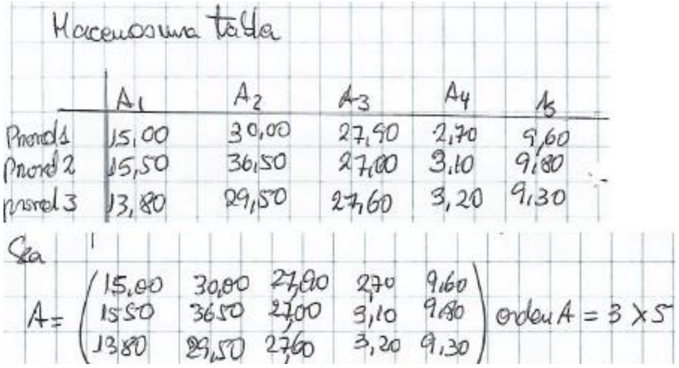
Students in class solve the problem posed. The teacher performs accompanying work, verifies that students play the assigned role, that they use the steps of the ABP as a guide that they actively participate by providing ideas for the solution of the problem.

After the job is complete, you must record it in a record that they are must upload to the wiki assigned to their group. The record should show that the first five steps of the PBL procedure have been developed.

The following describes what was worked by one of the groups. They queued the data in matrix form. They defined a price matrix; in the rows they placed the vendor data and in the columns the data of the item types (Fig. 1). Similarly, they defined an order matrix (Fig. 2) in which the data per month and the number of items of each type were placed in the rows.

TABLE I. ORDERS MADE

Month	Notebooks	Pens box	Ink cartridges	Staples boxes	Bond paper
Oct.	11	5	5	4	14
Nov.	12	1	8	8	18
Dec.	15	8	10	12	20



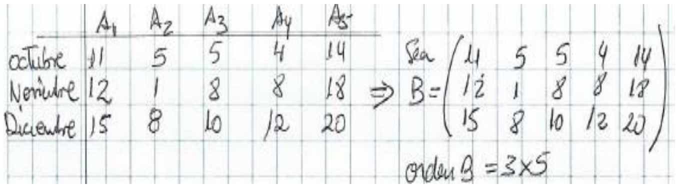
Hacemos una tabla

	A_1	A_2	A_3	A_4	A_5
Provd 1	15,00	30,00	27,90	2,70	9,60
Provd 2	15,50	36,50	27,00	3,10	9,80
provd 3	13,80	29,50	27,60	3,20	9,30

Sea

$$A = \begin{pmatrix} 15,00 & 30,00 & 27,90 & 2,70 & 9,60 \\ 15,50 & 36,50 & 27,00 & 3,10 & 9,80 \\ 13,80 & 29,50 & 27,60 & 3,20 & 9,30 \end{pmatrix} \quad \text{orden } A = 3 \times 5$$

Fig 1. Price Matrix.



	A_1	A_2	A_3	A_4	A_5
octubre	11	5	5	4	14
Noviembre	12	1	8	8	18
Diciembre	15	8	10	12	20

$$\Rightarrow \text{Sea } B = \begin{pmatrix} 11 & 5 & 5 & 4 & 14 \\ 12 & 1 & 8 & 8 & 18 \\ 15 & 8 & 10 & 12 & 20 \end{pmatrix}$$

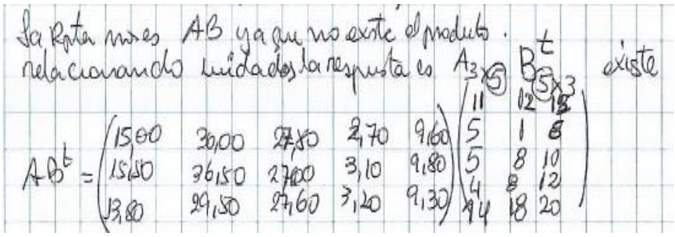
$$\text{orden } B = 3 \times 5$$

Fig 2. Data per Month and Prices Matrix.

To calculate the cost of those orders, with each vendor, they deduced that they would have to do a matrix multiplication. They had to check the compatibility of the dimensions, it had to be true that the number of columns in the first array was equal to the number of rows in the second. They needed to reason on which of the arrays should be reordered for the units to be consistent (matching prices and item numbers). Therefore, the product they made is shown in Fig. 3

Fig. 4 shows the results of the operation. The group concludes that it must be purchased from the third supplier.

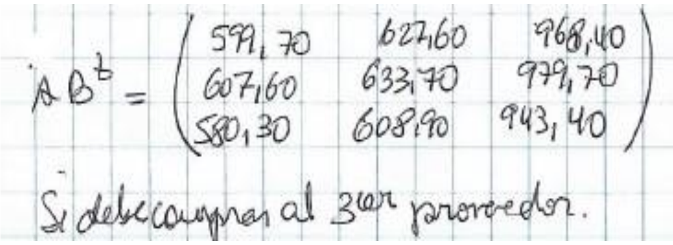
Since the behavior of the groups was relatively homogeneous, students were asked to proceed in a similar manner with the other proposed problems. Each proposed problem dealt with a particular type of operations between arrays.



la Rpta no es AB ya que no existe el producto.
relacionando unidades la respuesta es $A_{3 \times 5}\, B^t_{5 \times 3}$ existe

$$AB^t = \begin{pmatrix} 15,00 & 30,00 & 27,90 & 2,70 & 9,60 \\ 15,50 & 36,50 & 27,00 & 3,10 & 9,80 \\ 13,80 & 29,50 & 27,60 & 3,20 & 9,30 \end{pmatrix} \begin{pmatrix} 11 & 12 & 15 \\ 5 & 1 & 8 \\ 5 & 8 & 10 \\ 4 & 8 & 12 \\ 14 & 18 & 20 \end{pmatrix}$$

Fig 3. Matrix Product.



$$AB^t = \begin{pmatrix} 599,70 & 627,60 & 968,40 \\ 607,60 & 633,70 & 979,70 \\ 580,30 & 608,90 & 943,40 \end{pmatrix}$$

Si debe comprar al 3er proveedor.

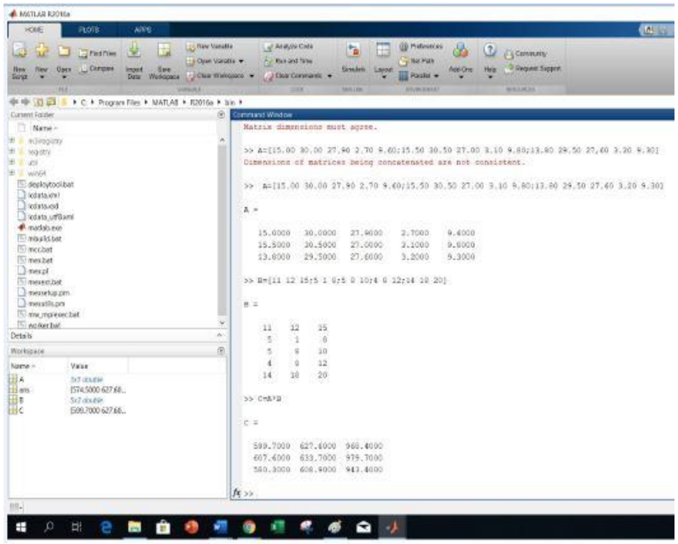
Fig 4. The Matrix Product Results.

3) Stage 3: The final report delivery

Here the suggested final step for the PBL methodology was developed. For the resolution of the other proposed problems, a face-to-face session was combined with autonomous out-of-class work.

The proposed problems related to different themes: graphs, rotation matrices and mixtures; consequently, the members of each group had to collect information from different sources: web browsers, literature of the subject, among others.

Partial progress was uploaded to each group's wiki. The process was followed up and feedback was given on the progress presented.



```
Matrix dimensions must agree.

>> A=[15.00 30.00 27.90 2.70 9.60;15.50 30.50 27.00 3.10 9.80;13.80 29.50 27,60 3.20 9.30]
Dimensions of matrices being concatenated are not consistent.

>>  A=[15.00 30.00 27.90 2.70 9.60;15.50 30.50 27.00 3.10 9.80;13.80 29.50 27.60 3.20 9.30]

A =

   15.0000   30.0000   27.9000    2.7000    9.6000
   15.5000   30.5000   27.0000    3.1000    9.8000
   13.8000   29.5000   27.6000    3.2000    9.3000

>> B=[11 12 15;5 1 8;5 8 10;4 6 12;14 18 20]

B =

    11    12    15
     5     1     8
     5     8    10
     4     6    12
    14    18    20

>> C=A*B

C =

  599.7000  627.4000  968.4000
  607.4000  633.7000  979.7000
  580.3000  606.9000  943.4000

fx >>
```

Fig 5. MatLab Check of Results of the Matrix Product.

Finally, each group produced a final report in which, taking into account the feedback received, the process followed and the results they had reached were allocated. Additionally, for the final report it was requested to check the results using MatLab. Fig. 5 shows the solution in MatLab for the problem described.

4) Stage 4: The final report presentation

A plenary session was scheduled for submission. In order for students to acquire practice in reporting in limited-time formats, they were asked to search for information about the "pecha kucha" format.

The plenary was held by assigning each group 7 minutes for the exhibition and 5 questions.

As there was the feeling that time had been very short, in the virtual classroom a P and R forum (questions and answers) was opened that is characterized because to participate it is necessary to issue an opinion first. Students were informed that everyone should participate at least once in this forum.

5) Stage 5: Activity evaluation

The evaluation of PBL's activity considered two aspects:

- Evaluation of group work: The final report was evaluated using a rubric (Table II, at the end of the document).
- Evaluation of individual learning: An evaluation was prepared through the "Questionnaire" functionality of the virtual classroom.

For the final qualification, weights were established for the different activities carried out: Process Evaluation (30%), Final Report (30%), Individual Learning (40%).

TABLE II. RUBRIC TO EVALUATE PBL ACTIVITY

Evaluation criteria	Higher (from 18 to 20 points)	High (from 15 to 17 points)	Basic (from 11 to 14 points)	Low (less than 11 points)
Obtaining information	Search, organize and select information in a relevant way	Identify important information, but omit some aspects.	Locate the information, but it is difficult to relate the different facts.	Search for the information, but cannot contextualize it to solve the problem
Understanding	Identify the main idea and determine its purpose	Find the main idea but do not synthesize the information.	It does not use all the relevant information.	It does not identify the main ideas.
Interpretation	Contrast the information and make inferences	Analyzes the information, but fails to relate aspects that lead to a better interpretation.	Analyze parts of the information, but does not identify specific aspects	It makes no sense to the information, then does not interpret it properly.
Reflection and assessment.	Relate the information to the problem.	He/she needs to organize his ideas better and relate them to the problem	It stays on the literal level and does not relate it to the problem.	Read the problem, but do not relate the information to a possible solution.

VI. Results and Discussion

The results of the survey conducted in order to gather the perception of students, in relation to the experience of Problem-Based Learning, are shown in Fig. 6 (at the end of the text).

For the conduct of the survey, a questionnaire was designed that considered eight aspects that were valued using a three-tier Likert scale.

In the light of the results obtained, it is confirmed that the use of the methodology "Problem-Based Learning" enables students to assimilate and transfer concepts and develop thought strategies.

In the design of a problem-solving-based teaching sequence, it is not enough to apply the steps suggested by the methodology; certain fundamental objectives should be taken into account: (a) the assimilation of concepts and principles, (b) the ability to transfer to real problems, (c) the development of analysis and synthesis capabilities and (d) the development of strategies for the solution of Problems.

Applying a problem-solving methodology, supported by ICT tools, can favor some important aspects of the teaching-learning processes of Mathematic: student motivation, development and creativity, development of mathematical thinking and the achievement of meaningful learnings.

The study showed that students positively value the PBL methodology, 83% of them perceive that they favor their learning and about 79% agree with [11, 12] that the use of this allows the development of generic competencies such as reflection creativity and interpersonal communication.

In relation to the time allotted for the activity, only 45% of students perceive that the time was sufficient, which is explained by their lack of experience in working as a team; as students of the first university cycle, they bring with them the idea that working as a group means the non-systematic union of individual efforts.

On average, 60% agreed that good integration between theory and practice was verified, while around 65% agree that they were pleased with the methodology; however, only 52% stated that the methodology was accessible. This could be explained through resistance to the change in teaching methodology, a situation already discussed in [13].

Overall, the qualitative results have been quite satisfactory, results that match [14], experience in which the ABP methodology has been applied to enhance students' learning outcomes.

On the other hand, students suggest more activities where they can view the application of the contents.

Considering that students take their first year and the nature of the subject, it is difficult to find applications of some subjects, however, it is a challenging task that will try to fulfill for the next course.

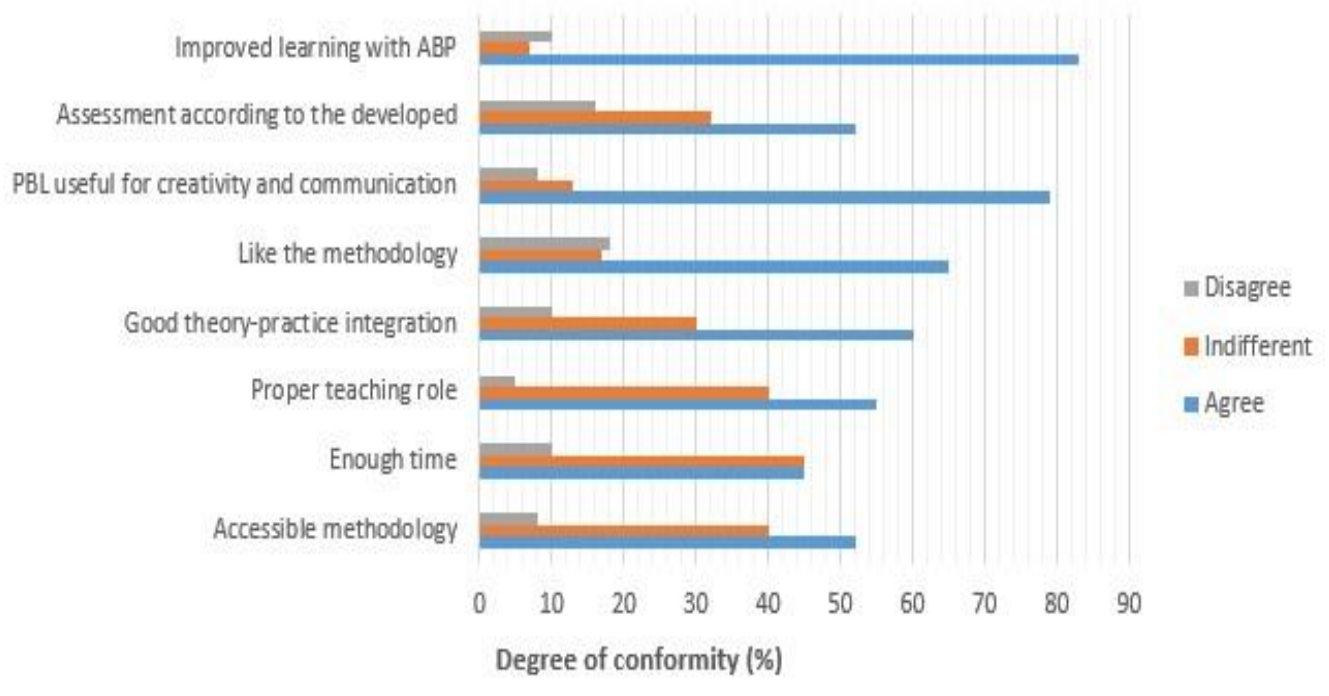


Fig 6. Opinion Survey Results.

VIII. Conclusions

During the interaction process, students in addition to understanding and solving the problems recognized the applicability of the object of study and the need to work cooperatively.

Implementing an ABP activity to solve array operations application issues has helped students become familiar with various applications.

The results indicate a high degree of satisfaction with the work done and commitment to your learning process; which shows the need to change traditional methodologies in the teaching of Mathematics (professor-centered) to active methodologies (student-centered).

Problem-Based Learning is an alternative to the traditional methodology, which allows to evaluate the student's ability to solve situations comparable to real situations, bringing them closer to the professional profiles required by today's society.

Acknowledgment

We want to express the authors' our sincere gratitude to the National University of San Agustín de Arequipa for the support received for the realization of the proposal and we hope that the results will benefit the institution.

References

[1] J. Kaput. “Transforming algebra from an engine of inequity to an engine of mathematical power by algebrafying the K-12 curriculum”. Dartmouth, MA: National Center for Improving Student Learning and Achievement in Mathematics and Science. 2000.

[2] C. Kieran. “Algebraic Thinking in the Early Grades: What Is It?” The Mathematics Educator, 18(1), 139-151. 2004.

[3] C. Cowen. “A project on circles in space”. En D. Carlson, C. R. Jonson, D. C. Lay, A. D. Porter, A. Watkins y W. Watkins (comps.) Resources for the teaching of linear algebra (pp. 59-70). Washington, Estados Unidos: Mathematical Association of America. 1997.

[4] J. Day. “Teaching linear algebra new ways”. En D. Carlson, C. R. Jonson, D. C. Lay, A. D. Porter, A. Watkins y W. Watkins (comps.) Resources for the teaching of linear algebra (pp. 71-83). Washington, Estados Unidos: Mathematical Association of America. 1997.

[5] A. García. “Aprendizaje basado en problemas: aplicaciones a la didáctica de las ciencias sociales en la formación superior”. Ponencia en el II Congrés Internacional de DIDACTIQUES. Girona, Francia. 2010.

[6] N. Bedregal, “Cooperative learning using Moodle as a support resource: Proposal for continuous evaluation in operational research”, Proceedings - International Conference of the Chilean Computer Science Society, SCCC. Volume 2017-October, 5 July 2018. DOI: 10.1109/SCCC.2017.8405131.

[7] N. Bedregal-Alpaca, D. Tupacyupanqui-Jaén AND V. Cornejo-Aparicio. “Video and Cooperative Work as Didactic Strategies to Enrich Learning and Development of Generic Competences in numerical Methods”. 2018 XIII Latin American Conference on Learning Technologies (LACLO). 2018. DOI: 10.1109/laclo.2018.00038

[8] A. Romero y J. García, “La elaboración de problemas ABP”, El aprendizaje basado en problemas en la enseñanza universitaria (pp. 37-53). Murcia: Editum, Ediciones de la Universidad de Murcia, 2008.

[9] N. Bedregal, N. y D. Tupacyupanqui, “Integration of active methodologies and virtual classroom in the teaching-learning processes of Discrete Mathematics”, Proceedings of the LACCEI international Multi-conference for Engineering, Education and Technology Volume 2018-July, 2018. DOI: 10.18687/LACCEI2018.1.1.81

[10] N. Bedregal-Alpaca. V. Cornejo-Aparicio, D. Tupacyupanqui-Jaén and S. Flores-Silva. “Evaluation of the student perception in relation to the use of the Moodle platform from the TAM perspective”. Ingeniare. Rev. chil. ing. vol.27 no.4 Arica dic. 2019. DOI 10.4067/S0718-33052019000400707.

[11] C. Calpopiña and S. Bassante. “Aprendizaje basado en problemas, un análisis crítico”. Rev Publicando. 2016;3:341-50.

[12] N. Bedregal-Alpaca, D. Tupacyupanqui-Jaen, M. Rodriguez-Quiroz, L. Delgado-Barra, K. Guevara-Puente and O. Sharhorodoska, "Problem-Based Learning with ICT Support: An experience in teaching-learning the concept of derivative," 2019 38th International Conference of the Chilean Computer Science Society (SCCC), Concepcion, Chile, 2019, pp. 1-7. DOI: 10.1109/SCCC49216.2019.8966396

[13] J. Vadillo, , I. Usandizaga, A. Goñi and M. Blanco. “Análisis de los resultados de la implantación ABP en un Grado de Ingeniería Informática”. Actas del simposio-taller sobre estrategias y herramientas para el aprendizaje y la evaluación. 2015.

[14] L. Sarkady, L. Alveiro, M. Carrasco, M. Molina, M. Llanes, M. and M. Aguado. “Investigaciones educativas sobre enseñanza y aprendizaje de la Química”. 30.o Congreso Argentino de Química. 2014

Project based Learning Application Experience in Engineering Courses: Database Case in the Professional Career of Systems Engineering

César Baluarte-Araya
Professional School of Systems Engineering
Universidad Nacional de San Agustín de Arequipa
Arequipa, Perú

Abstract—**In many universities, training research is applied in courses as a basic element of research and fundamental in the professional training of every student, which result in strengthening and increasing knowledge about certain areas, as well as to achieve skills, competence, abilities and attitudes. The present work shows the formal application experience of the Project Based Learning (ABPr) methodology in Database Course (BD) at the Professional School of Systems Engineering (EPIS) of the National University of San Agustín (UNSA), Arequipa-Peru, accommodating the nature of the course being the theory taught by a teacher and laboratory practices by another teacher. The goal is to apply an active teaching strategy to an engineering training course. The methodology used is Project-Based Learning for a research project training for a real problem in an organization to be developed by each team in the semester, with deliverables that will be evaluated by grade scale and the formative research report assessed through the rubric; the input and feedback that the teacher makes of them serves for the improvement and experience in the training of the student. The results obtained show that the objectives in the training of students were achieved, as well as the development of the competencies related to the course, in addition that the application of ABPr gives good results for courses of an engineering career serving as feedback for the continuous improvement of this course and experience for the implementation of ABPr in other curriculum courses. Concluding that formative research as a pillar of a basic level of research initiation is given in a cross-cutting way in the curriculum courses, that the active teaching strategies properly planned and properly applied to each reality allow to achieve the desired results such as: increase knowledge of the area, strengthen skills, abilities, attitudes as the case of the present.**

Keywords—Project based learning; formative research; competences; skills; formative research report

I. Introduction

Student training at the university level, particularly in the area of engineering, has in recent years been treated in addition to conforming to international standards such as [1] and other Spanish-speaking standards such as [2],[3] which accredit curricula university is to include in courses of curricular plans the formative research that aims to enable graduates to achieve in their vocational training to base an adequate level of research within their skills to develop knowledge and apply it to change the reality for the good of the society to which they are due; thus, the teacher by including in the course syllable the training research can use teaching strategies [4], [5] appropriate to the nature of the course to achieve the objectives and competences.

With previous experience in the Professional School of Systems Engineering [6], of the National University of San Agustín [7], Arequipa – Peru, in the course of Writing Articles and Research Reports (RAII) taught in the III semester since the academic year 2014 in which the preparation of poster and documentary research article is available as deliverables, thus having a first level of formative research undergoing specialty and of developing in the student soft skills, abilities; the development of other skills and competencies needed to achieve a higher level of research is determined in the database course taught in the fifth semester of the Curriculum to apply the Project-Based Learning (ABPr) methodology for the database design and as another result the preparation of the Formative Research Report; that admits in some way its evaluation and validate what has been done, and to strengthen soft skills and abilities in the use of techniques and tools.

UNSA decided to include in the syllable structure of the courses of the Curriculum Plan of the Professional Schools training research to address the shortcomings of students in knowledge, skills and abilities, making the teacher use the best teaching strategies in teaching learning the institutional educational model by competencies; is so for the Database course the Project-Based Learning methodology is applied to the project of designing a database of a real problem of an organization and that through the software developed using the methodologies, techniques and tools the student manages to validate the elaborate design; delivering the training research report based on the template used that includes items such as: executive summary, introduction, objectives, theoretical framework, methodology, results, conclusions, recommendations, references, annexes, self-assessment; which will be evaluated with the relevant rubric.

The objective pursued is to apply the ABPr for the first time formally to the System Engineering Professional Career Database course in solving various real-world problems by the trained work teams; managing to produce reports by the deliverables of established laboratory sessions, the training research report, to increase their skills and abilities to propose solutions in changing environments, interact in situations, take decisions, work as a team; as it refers [8] in this new knowledge society.

The research is descriptive level and the methodology used for development is based on the phases of problem solving, of the scientific method [9] who states that the student allows to develop skills; for the present work adapting and expanding others; the survey technique and its instrument are also applied to the questionnaire to obtain the data, systematize it, analyze the results to reach the relevant conclusions pursued by the research.

Achieving as results that the students in their working group applying the ABPr developed the project of designing a BD of a problem of reality, which to be validated develops the software application thus validating the design by complying with the user requirements and have the necessary functionality. Thus also students have the perception and recognize having applied and achieved perfecting their skills, abilities, attitudes and valuing the result achieved in the project worked in the course.

Conclusions are reached such as that the Project-Based Learning methodology is ideal for enhancing autonomous learning, developing skills, abilities and achieving good performance in students; recognizing that what has been achieved in the real-life problem project serves for their future professional performance in the sectors where it will perform; it also reinforces the process of research training which takes place in a cross-cutting way in this course and of being able to take place in others of the Curriculum Plan.

The paper is organized in following sections; Section II of Related Works – Context of Experience, Section III deals with the Design of the Formative Research Project, Section IV shows the Method of Work that shows the stages for the development of the project, Section V shows the perceptions that students have of the ABPr strategy and of the Experiences left by the Development of the Database course with this strategy, Section VI shows the results of the Evaluation of the Research Report Formative, Section VII shows the Lessons Learned from Development and that leads to continuous improvement, in Section VIII the Discussion that is made is touched, that others investigated or worked and the result obtained that contributes to the integral training of the student , Section IX presents the conclusions of the work, Section X shows the future work that could be enhanced, Section XI shows the respective appreciation, and Section XII shows the respective References.

II. Related Works: Context of Experience

A. *Formative Research*

In this topic [10] addresses how the problem of the teaching-research relationship contemplates the role that research can play in learning the same research and knowledge. Thus he considers that "formative research" has to do with the concept of "training", of shaping, of structuring something throughout a process. This training, for our case, refers to students who are prepared, through the activities developed, to understand and advance scientific research in their training.

It focuses on the learning strategy by discovery and construction of knowledge, where the student is the center of the process, starting from a problem seeks, finds, examines sources of environments or similar success stories, related writings, collecting the data, organizing it, classifying it, ordering it, interpreting the results until we get statements of proposals or solutions. Interesting to mention that teachers use methods, practices for formative research with students from undergraduate work, theoretical essay, in research work with the teacher when working rigorously in advising research. In some way it refers [11] that learning is participatory, student-centered and the use of active methodologies used in the area of accounting by quoting Activity-Based Learning and Project-Based Learning.

Thus also [12] contemplates in the sub-process: Management of Formative Research where they establish mechanisms to integrate the research process into the teaching-learning process; where the task of Developing Didactic Strategies in the teaching-learning process designs the methodological mechanisms where problem-based learning, case study, portfolio, without neglecting others such as learning based on projects would narrow down as well.

For his part [13] in his study he analyzes the role of formative research in the development of undergraduate scientific competences by driving scenarios that drive research processes; and to assess the perception of students of their contribution and impact of the vocational training process on the development of their research skills; showing the quantitative analysis of the at-attitudinal profile resulting from the main indicators on the scale of attitudes presented: There should be more incentives and incentives to promote research with 69.2 points; I believe that research is a very important tool for generating new knowledge with 62.5 points; I think my program should re-evaluate how it guides the research with 50.0 points.

Since 2014, exists the first experience in EPIS to initiate the formative research reflected in the work of [14] with reference from results of previous works [15-16] in demonstrating and validating the achievement of students in developing articles of and in addition to reaching skills and developing soft skills.

B. *Project based Learning*

As it was discussed in [14] referring to [1] that says ..."Our future graduates should be able to work in a global environment in multidisciplinary teams, solving constantly changing problems. In addition to much of what students have learned (technologies, software, methodologies, programming languages, etc.) will have changed by the time they graduate."; there is an expectation that students will be warmed up in addition to the technical knowledge that they are from the curriculum, with soft and professional skills in their training as engineers.

For the [17] ABPr, “It is a learning methodology in which students are asked, in small groups, to plan, create and evaluate a project that responds to the needs posed in a given situation.”

There are many experiences of universities in the world that use active methodologies, such as to cite Spanish, [10], [11], [12] where the focus is on the student, acquisition of new knowledge, participatory teamwork and collaborative, concrete results of the problem of reality, so also for [18] is learning to solve complex problems of applying to real situations the knowledge and skills acquired in their training through the project planning, designing, carrying out a set of activities in a given time; for [19] ABPr is relevant to students' work in learning essential knowledge and skills in a real project in real-world contexts in the Software Project Workshop engineering course, where learning is directed to the skills development and evaluation is with respect to performance in effective work. However, some teachers employ a scenario of a fictional project that can take place in the professional exercise [20] where the student is given the information of the organization showing problems and other aspects, enhancing the autonomous and increase academic performance; collecting by survey opinions on the methodology and academic results of the course.

The ABPr has characteristics that we will consider summarizing those of [4]:

- Affinity with real situations, in the world of work
- Practical relevance, theoretical-practical exercise of job insertion and professional development
- Student-oriented approach to their interests and needs
- Action-oriented approach, autonomous concrete actions, both intellectual and practical
- Product-oriented approach, obtaining relevant results subject to valuation and criticism
- Process-oriented approach, Learn to learn, learn how to do and learn to act.
- Self-organization: Goal determination, planning, realization and control are decided and carried out by the students themselves.
- Collective realization: Students learn and work together in the realization and development of the project.
- Interdisciplinary character: Through the realization of the project, different areas of knowledge, subjects and specialties can be combined.

C. Learning Assessment

The assessment of learning when applying ABPr in the teaching learning process of the subject treated can be carried out through instruments that adapt to each reality studied and to be resolved, for the case the rubric [21-23] is used through the establishment of criteria and levels of achievement to assess knowledge, learning achieved, achievement of skills, using a scale of qualification, making students known at the beginning of the course their content and description of rubric.

It is important to feedback from students' learning process as they worked [23] by comparing feedback to evaluate with rubrics to their experimental group to see if it facilitated self-reflection, self-assessment by students and improve learning outcomes versus traditional feedback; reached conclusions where feedback is effective if it focuses on learning and makes it easier for the student to develop self-regulation skills, self-sensing and found significant differences in all forms of assessment in the experimental group. Also finding differences between the different forms of evaluation and the learning results obtained.

If constructive guidance is taken into account in the curricular proposals [24] consider the relationship between evaluation methods using as criteria the desired effects of the evaluation, reaching the current didactic syllalists adapting to reality or nature of each course and the activities that are programmed to achieve learning; with the intention of giving students the opportunity to continue learning, this requires an evaluation as part of an order generating learning experiences for the actors.

In addition [25] from the competency approach, it sees the evaluation aimed at evaluating students' performance against an activity or problem of their profession being referenced in evidence and indicators, this objective and consistent assessment of the activities; also allowing to evaluate the competencies related to the application, synthesis, criticism reflecting the lesser or greater mastery of competence on the part of the student.

III. Design of the Formative Research Project

For the development of the project, aspects were considered to be relevant and to deal with a problem of reality where the student expands the knowledge and applies the methodologies, techniques and tools related to the course of his curricular plan, which in addition to research activities allow the student to produce reports with defined structure and be evaluated.

Database is a practical theoretical course developed in the shared chair of a professor in theory and a professor in laboratory practices. The work teams develop the design project of a BD and developing the software application as a result that validate design by meeting requirements and having functionality.

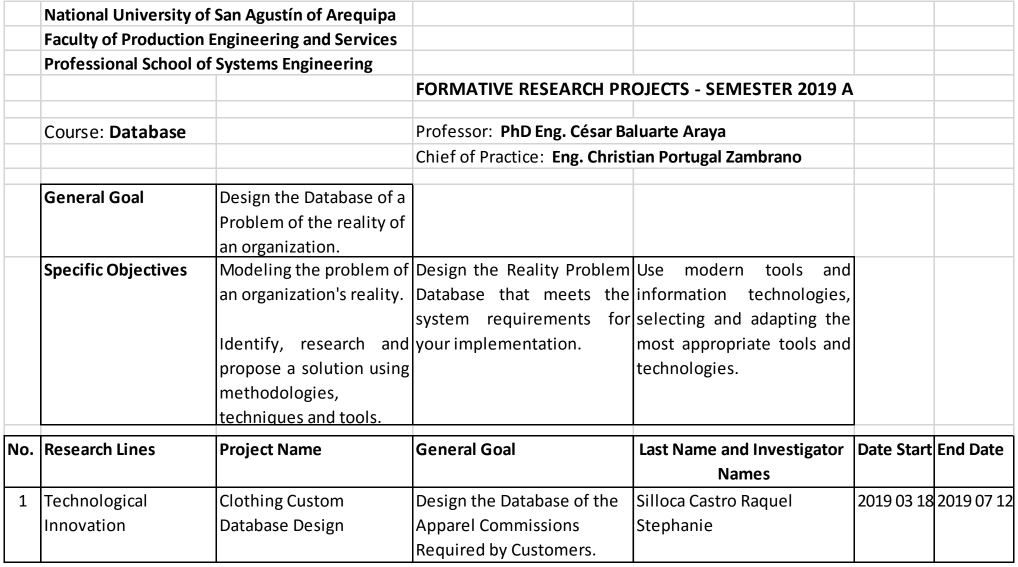
For the case there are scenarios of the various real organizations both business and governmental, where the student assumes the specific technical roles in the organization of the project, having to relieve the data and information of the problem main subject of the project.

A. The Project

The project to be developed has as features:

- Be a problem of a real organization
- The team of 2 students
- Each team's projects will be recorded
- Will be developed throughout the academic semester.

Fig. 1 shows the record of the formative research projects of the course.



	National University of San Agustín of Arequipa					
	Faculty of Production Engineering and Services					
	Professional School of Systems Engineering					
			FORMATIVE RESEARCH PROJECTS - SEMESTER 2019 A			
	Course: **Database**		Professor: **PhD Eng. César Baluarte Araya**			
			Chief of Practice: **Eng. Christian Portugal Zambrano**			
	General Goal	Design the Database of a Problem of the reality of an organization.				
	Specific Objectives	Modeling the problem of an organization's reality. Identify, research and propose a solution using methodologies, techniques and tools.	Design the Reality Problem Database that meets the system requirements for your implementation.	Use modern tools and information technologies, selecting and adapting the most appropriate tools and technologies.		

No.	**Research Lines**	**Project Name**	**General Goal**	**Last Name and Investigator Names**	**Date Start**	**End Date**
1	Technological Innovation	Clothing Custom Database Design	Design the Database of the Apparel Commissions Required by Customers.	Silloca Castro Raquel Stephanie	2019 03 18	2019 07 12

Fig 1. Formative Research Projects. Source: Own Elaboration.

B. *Project Theme Assignment*

In the development of the laboratory sessions of the database course from 2014 the sessions of a project to be developed during the semester by groups consisting of 2 students, according to the considerations for this: disposition of communication, meeting, working hours for the project and study of other courses of the team members, compatibility of personal interrelationship; giving very good results. Having that precedent in this experience it was decided to continue with the assignment of random topics to the trained groups, who go to the organizations to request to carry out the project of the assigned topic, considering that if in the organization they pose another issue of a problem to work they can do so.

C. *Tracking*

As the semester progresses, the feedback of the results of the deliverable is made in each laboratory session, which leads the student to make the relevant modifications that will serve it for the following future sessions due to the nature of the project in the course.

The teacher plays the role of counselor by giving suggestions, guiding and helping the team, students develop the project as independently as possible.

D. *Evaluation*

The evaluation at the laboratory sessions from 2008 to 2017 was considered to be carried out of the respective deliverable according to the checklist criterion established for adequate compliance with the result to be delivered.

For the year 2019 in the new competency curriculum the evaluation for the engineering area is adopted as the main instrument for evaluating the course's competencies; the topic's rating scale or development checklist for lab deliverables, which are stored in a virtual classroom repository resembling a portfolio by deliverable, is used in this course; and the rubric for the evaluation of the final formative research report.

The theoretical development of the course was determined to use the rubric evaluation strategy that makes it better to achieve a better result and achieve the objectives. What is stated by [21] is a good criterion of carrying out the learning and evaluation tasks in the student-teacher relationship, and by [23] that makes it easier for the student to be consistent with how far his learnings go and what is the maximum desirable level; in order to achieve:

- Set achievable goals, conducting a Real BD Design, an application that validates the BD, a formative research report.
- Develop a project using ABPr based on a methodology established for the case.
- Carry on the part of the teacher monitoring and feedback through the deliverables and their assessments for continuous improvement.
- Increase the interpersonal and social relationships of the participants of the task team.
- Manage the time for the execution of deliverables.

IV. Method of Work

A. *Conceptual Design*

There are for the development of the project experimental studies such as that of [08] [19] [20] that serve as a reference and generate something proper to the course of BD at the university level in the professional career of Systems Engineering.

The purpose is for the student to search, research, review similar situations, review related literature, collect, organize, interpret data and list alternatives for solving a real problem of reality in an organization, but then the ABPr strategy that targets the BD course is used to strengthen and discover greater knowledge and also the development of skills, abilities, and assessment of its results.

The research developed is of descriptive level and the methodology used for development is based on the phases of problem solving, based on the scientific method, which allows the student to develop skills such as: delimit a problem, formulate solution hypotheses, design experiments, observe, measure, gather information and data, analyze them, draw conclusions; for the present work adapting and expanding others; the survey technique and its instrument of the questionnaire are also applied to obtain the perception data, systematize it, analyze the results to reach the relevant conclusions pursued by the research, and be taken into account for continuous improvement.

The stages defined for project development are shown below:

1) Starting point

Main theme

- Design and implementation of a database
- Initial Question
- With the knowledge gained and complemented by the database, it will be possible to design and implement a database in a real-world situation?

2) Collaborative Team Training

- Team / group of 2 people.

3) Definition of the Final Product

Product to be developed.

- Design and implement the database of a module or subsystem of information of a real problem of a real organization.

What to know (learning objectives)

- Perform conceptual modeling of the database
- Perform the logical design of the database
- Perform the physical design of the database
- Perform application development to validate the database.

4) Organization and Planning

Role assignment

- Team members' roles will be rotating for each evaluable session, considering:
    - Project Manager
    - Project Analyst
    - Application Scheduler
    - Defining tasks and times
- Planning
- Coordination
- Control.

5) Search and Collection of Information - in the Organization

- Review of objectives.
- Recovery of Previous Knowledge, course material, bibliography, references in research database (Scopus, Web of Science, Google Academics, etc.).
- Introduction of new concepts.
- Search for information (primary sources, research thesis, success stories, other sources).

6) Analysis and Synthesis

- Sharing, sharing information; organization of information and prepare it using tools appropriate to the project.
- Contrast of ideas, debate; analysis of the situation of the problem to be solved.
- Troubleshooting; according to the stages of database development in organizations.
- Decision-making; to perform the following activities.

7) Design

- Conceptual Design
- Logical Design
- Physical design.

8) Development of the Application

Demonstrating the proper design of the database; using development tools, and demonstrating their applicability in solving the problem.

9) Project Presentation

- Present the results of the project or its stages
- Review and Evaluation of the project or its stages.

10) Preparation of the formative research report

Prepare the report in accordance with the drafting template.

11) Evaluation of the formative research report

- The teacher evaluates the report using the rubric.
- Includes general self-assessment of the work team.

Techniques and Instruments to Use

A. Techniques and Instruments

- Project-Based Learning
    - Project development – project of the course, – in stages within the development of the course in database laboratories.
- Survey
    - Questionnaire – appreciation on the part of students.

- Project Development – course project, which will be completed in stages during the development of laboratory practices, which are evaluated.

B. Evidence

- Digital files, lab work.
- Moodle, as a repository for course work.
- Portfolio, which will initially be assembled from the project outline, laboratory guides, data dictionary, the work performed considered as deliverables.

Tools

The tools to use are:

- Computer
- Software
- Operating System; Linux, Windows or other
    - Programming language; Java, PHP or other
    - Database Management System, MySQL, Postgres, or other.
- Application tools
    - Suite Office
    - Data Dictionary
    - TOAD.
- Other
    - Standards for BD Project - Development - Offprint UNSA 2019.

B. Participants

The BD course of the Systems Engineering career is taught in the fifth semester (3rd year), has 6 hours per week, 4 theoretical and 2 in laboratory, the semester has 17 weeks; participating in the evaluation of the Formative Research Report 74 students being the sample size; the theory of the course was in charge of a teacher and the laboratory practices in charge of another teacher forming 4 groups.

C. Data Analysis Technique

The data analysis technique was through the processing of the assessment of deliverables, the evaluation of the final formative research report, as well as the application of the questionnaire instrument to the sample to collect data from the survey of statistical, the data were systematized in the EXCEL electronic spreadsheet, analyzing the results obtained as statistics: averages and frequencies, charts, tables and tables.

D. Instruments

The use of the instruments is envisaged:

- Laboratory Guides
- Training research report template
- Evaluation rubrics
- BD course evaluation matrix.

E. Techniques

As evaluation techniques:

- The rubric
- The Grading Scale.

F. Deliverables and Assessments

From the laboratory sessions the deliverables were determined, which are developed and delivered by each team whose evaluation leads to a respective qualification and continuous and immediate feedback that favors the learning process and the skills development.

Table I shows the deliverables of the course lab sessions in the academic semester.

TABLE I. LIST OF DELIVERABLES FOR EVALUATION WITH ABPR METHODOLOGY

Table I	***Descriotión Deliverable***
Session 1	Take the actions to take a tour of the SGBD Access to visualize its functionality and to contemplate the concepts that are immersed in the. Carry out the complementary work of viewing the videos of the Access Basico 2010 Course. Designing a company's Orders BD by considering the tables and data given from: Customer, Seller, Zone, Order, Item. Prepare the deliverable report.
Sesion 2	Develop the Data Dictionary of a Module or Information Subsystem of a real company
Sesion 3	Model a database from the work or case study or topic assigned and worked on the practice of data dictionary using an automation tool to help model data to be performed. Developing the Relationship Entity Model.
Sesion 4	Develop the Conceptual Schema from the use of an automation tool to aid data modeling. Use the TOAD tool to model data.
Sesion 5	Developing data modeling using standardization.
Sesion 6	Work in the working environment of the MySQL database management system. Create the database of your lab job topic in the MySQL database management system. Establish the connection to the MySQL database to work with the JAVA programming language.
Sesion 7	Develop data upload programs to the MySQL database. Establish the connection to the database to then manipulate the data with the application programs.
Sesion 8	Manipulating data using the SQL functions of a MySQL database. Handling the added functions. Database View Management.
Sesion 9	Manipulating data using SQL triggers from a MySQL database. Handling stored procedures.

Source: Own elaboration.

The structure of the Formative Research report is shown below:

Cover
Index
Executive summary
I. Introduction
II. Objectives
1. Overall objective
2. Specific objectives
III. Theoretical Framework
IV. Methodology
V. Results
VI. Conclusions
VII. Recommendations
VIII. References (IEEE, APA)
Annexes
Self-assessment

V. Perceptions, Data base Course Development Experiencies

The following perceptions and experiences are presented as a result of applying ABPr in the BD course project:

A. *Perceptions*

- It is perceived that the student assimilates in each deliverable what at the beginning of the course had as knowledge, the objective that is pursued and to know what in the future will help in subsequent courses of the career.
- There is a perception that there is motivation to carry out a project of a real case or problem of an organization, which is systematically developed until the final product is obtained.
- Having feedback on the various partial deliverables by the review and observations made by the teacher for the correction, improvement and continuation of the project.
- The student is perceived to reinforce the skills and abilities by assessing the level achieved that allows to:
  - Properly draft the formative research report
  - Conduct adequate documentary research of the theoretical framework complementary to that provided to increase its knowledge.
- It is perceived that at the end of the course the student achieves the competencies.
- The objectives of the course have been achieved.

B. *Experiencies*

You have the following experiences of the developed course:

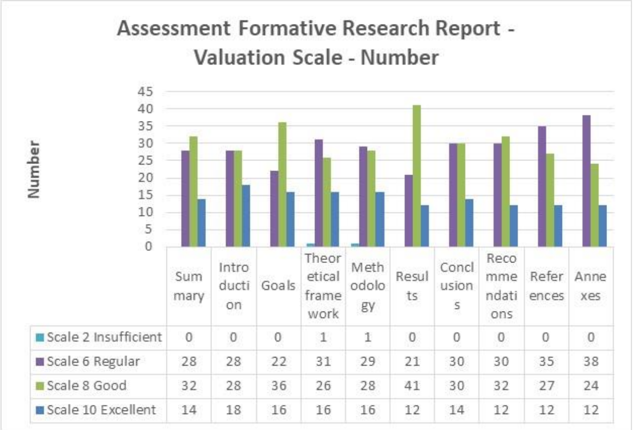
- For the first time, Project-Based Learning is formally and systematically used for an engineering specialty training course.
- Use the fundamental phases of scientific research with adaptation to Project-Based Learning that is reflected in the final report of formative research prepared by students.
- The skills and abilities achieved by the student in the RAII course are reinforced by being contemplated in the activities raised in the BD course project and through formative research.
- The time experienced by 38% of students having contact for the first time with reality in an organization like the company, which breaks the taboo of "what it will be".
- Designing practice guides that help students have a base support to develop the required deliverable.
- The use of the methodologies, techniques and tools necessary for the development of each project deliverable.
- Use the data dictionary as a fundamental tool for BD design as it supports the development of other project activities.

VI. Results

The results of the Evaluation of the Training Research Report (EIIF) carried out with the corresponding rubric of the BD course are shown below.

BD is a course where the student ultimately drafts a formative research report under a previously defined scheme and used for the purpose. This allows previously exposed to developing skills in writing, communication, among others.

Fig. 2 shows the number of students who were evaluated according to the criteria, their level and rating scale according to the established rubric.



	Summary	Introduction	Goals	Theoretical framework	Methodology	Results	Conclusions	Recommendations	References	Annexes
Scale 2 Insufficient	0	0	0	1	1	0	0	0	0	0
Scale 6 Regular	28	28	22	31	29	21	30	30	35	38
Scale 8 Good	32	28	36	26	28	41	30	32	27	24
Scale 10 Excellent	14	18	16	16	16	12	14	12	12	12

Fig 2. AFRR – Valuation Scale – Number. Source: Own elaboration.

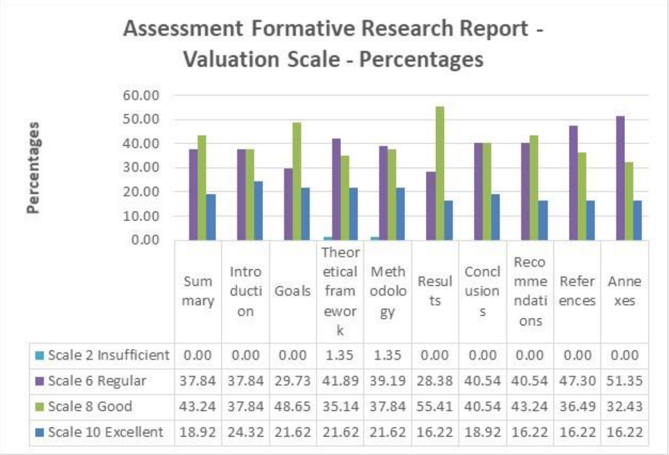
Fig. 3 shows the percentages of students who were evaluated under the same parameters, having the following interpretations:

- Students pass the assessment that is reflected from the regular level towards the excellent, which manifests to prepare the training research report in an appropriate way.

For the following interpretations it is considered to the Good and Excellent levels of the rating scale.

- The Summary criterion shows that 62.16% manage to develop it according to the structure normally used in its elaboration as: the topic to be investigated, objectives, results, conclusions and recommendations.
- By the Introduction criterion, 62.16% should clearly and punctually address the subject to be investigated, the objective, justification, benefits or contributions.
- The Objectives criterion has 70.27% to set and draft the overall objective and specific objectives in a clear and coherent way of the research topic.
- By the Theoretical Framework criterion it is given that 56.76% give a detailed description of each of the theoretical elements that were used directly in the development of research as: concepts, techniques, tools, research work related cases, success stories.
- The Methodology criterion shows that 59.46% reflect in addition to the type of research, method of information collection; contemplates the stages or phases used in the development of the project; clearly explaining each of them.
- The Results criterion has 71.62% including the results obtained from the project; to cite: Conceptual Design of the BD, Logical Design of the BD, Physical Design of the DB, Prototype of the software application that shows the implementation and operation of the DB.
- The Conclusions criterion shows that 59.46% draw clearly and in a timely manner the conclusions reached in the project that meet the objective set.
- The Recommendations criterion shows that 59.46% clearly and promptly draw the recommendations derived from the project results and based on the conclusions.
- By the References criterion, 52.70% have to draw up the list of references used in the report according to the specified standard or style.
- The Annexes criterion shows that 48.65% include other diagrams not included in the report body such as: tables, graphs, infographics, print screens; estimated expediently.

Based on the above, it can be determined that an average of 60.27% is achieved by students in an appropriate manner of the Formative Research Report; which is very promising as ABPr application experience for an engineering specialty course.



	Summary	Introduction	Goals	Theoretical framework	Methodology	Results	Conclusions	Recommendations	References	Annexes
■ Scale 2 Insufficient	0.00	0.00	0.00	1.35	1.35	0.00	0.00	0.00	0.00	0.00
■ Scale 6 Regular	37.84	37.84	29.73	41.89	39.19	28.38	40.54	40.54	47.30	51.35
■ Scale 8 Good	43.24	37.84	48.65	35.14	37.84	55.41	40.54	43.24	36.49	32.43
■ Scale 10 Excellent	18.92	24.32	21.62	21.62	21.62	16.22	18.92	16.22	16.22	16.22

Fig 3. AFRR – Valuation Scale – Percentages. Source: Own elaboration.

VII. Lessons Learned

The development of the ABPr database course leaves lessons learned that leads to continuous improvement resulting in the professional training of students, namely:

- Laboratory Practice Guides

It is confirmed that updating the topics allows the student to have a broad spectrum of information and knowledge to work on them.

- From the assessment

The rubric is used as an evaluation tool for both the theoretical and laboratory parts relating to the formative research report; which may have adjustments for the development of the next course.

VIII. Discussion

Based on the engineering concept of [8] ..."Engineering is the conceptualization, design, construction and management of projects and products aimed at solving a need of society or the environment". Thus the engineer must solve problems, for this it must have ingenuity, creativity, knowledge, skills, skills, analytical capacity, synthesis, and decision-making, among other aspects.

With the focus of skills training much has been written by authors such as INACAP [4], ITESCA [5] and considering that learning outcomes should be obtained using appropriate methodologies [18], from a revision of the ABPr model by [26] and [27] which reflects on the mention that it is being promoted as teaching methodologies in the work of projects, stating that it is understood by an appropriate context, as selected according to knowledge-related learning objectives to be built, conceived and applied a research process; which is also geared to incentivizing the culture of research as in many universities in Peru has been carried out in greater depth since the new University Law of 2014, based on it it has to be that many university educational institutions adopt changes that lead to a comprehensive training of their students; so through the norms of UNSA in research training based on transversal formative research in curriculum courses, formative research is taken in the syllables of the courses; taking, for example, that to date it has obtained results by having at the end of 2019, five graduates in the EPIS with the modality of scientific article published in indexed journal; being the first university at the Peru level recognized by SUNEDU (National Department of Education); this stems from the proper use of teaching strategies arising from social, economic, cultural, technological changes which require that the capital or human resource that is inserted into organizations be formed according to what the society sues, and how students carry out a project; BD case; in a certain time to solve a problem of reality in an organization or face an investigative task with proper planning, design, development, carrying out a set of activities, based on the progress and application of learning and effective use of resources.

This research work has succeeded in executing practical applications in the development of the BD course that assists in the integral training of the student in the field of knowledge, skills, abilities, reporting strategies, processing data and information, motivation to learn and achieve skills, achieving an adequate level of proficiency that allows for better development and advance in the levels of approval in your professional training; which also serves for job performance in the future.

IX. Conclusions

The following conclusions have been reached:

- Project-based Learning as an active methodology enhances autonomous learning by increasing knowledge, understanding topics, allowing to apply the knowledge acquired in class, allowing to strengthen the development of skills and abilities, as well as better academic performance, motivation and teamwork of students.
- The training process in formative research takes a cross-sectional course in the courses following the RAII course involved in the Curriculum Plan, as is the case with the Database course.
- Applying appropriate teaching strategies in other courses culminating in the thesis and thesis seminar essay project courses, allows to reach students with the highest level of research for undergraduate.
- Student recognition that what has been achieved in the development of the real life project in an organization serves for their professional performance in the sectors where it will take place in the future.
- Project development must be planned in great detail to achieve the desired results.
- Students reached the course competencies that contribute to their career competencies.
- The evaluation of the formative research report has made it possible to validate the achievements and gaps in the monitoring or tracing that is carried out to each task team that will allow the appropriate adjustments to be made for a better project development in the future course.

X. Future Works

- Conduct comparative research of the results obtained from applying ABPr in engineering specialty courses.
- Conduct a research of the tranverexited of applying active teaching strategies in the training of the student.
- Reiterate conduct training research in other courses for the development of other skills, abilities or competencies in students of the professional career; to complement their vocational training.

Acknowledgement

Our thanks to the National University of San Agustín of Arequipa for allowing and supporting the development of Formative Research that is of the interest of students.

References

[1] ABET. Why ABET Accreditation Matters. https://www.abet.org /accreditation/what-is-accreditation/why-abet-accreditation-matters/. Ultimo acceso diciembre 2020.

[2] ANECA, Agencia Nacional de Evaluación de la Calidad y Acreditación, España. www.aneca.es

[3] CNA, Consejo Nacional de Acreditación, Colombia. https://www.cna.gov.co/1741/articles-186359_pregrado_2013.pdf

[4] Subdirección de Currículum y Evaluación, Dirección de Desarrollo Académico, Vicerrectoría Académica de Pregrado, Universidad Tecnológica de Chile INACAP. (2017). Manual de Estrategias Didácticas: Orientaciones para su selección. Santiago, Chile: Ediciones INACAP.

[5] Rodriguez Cruz, Reyna, Compendio de estrategias bajo el enfoque por competencias, Instituto Tecnológico de Sonora ITESCA, México, 2007. http://www.itesca.edu.mx/documentos/desarrollo_academico/compendio_de_estrategias_didacticas.pdf.

[6] Escuela Profesional de Ingeniería de Sistemas. http://www. episunsa.edu.pe.

[7] Universidad Nacional de San Agustín de Arequipa. http://www .unsa.edu.pe.

[8] Rodríguez-Sandoval, Eduardo; Vargas-Solano, Édgar Mauricio; Luna-Cortés, Janeth, Evaluación de la estrategia "aprendizaje basado en proyectos", Educación y Educadores, vol. 13, núm. 1, abril, 2010, pp. 13-25 Universidad de La Sabana, Cundinamarca, Colombia.

[9] GARZA-RIVERA, RG. El rol de la física en la formación del ingeniero. Ingenierías, 2001, vol. IV, No. 13, pp. 48-54.

[10] Restrepo Gómez, Bernardo, INVESTIGACIÓN FORMATIVA E INVESTIGACIÓN PRODUCTIVA DE CONOCIMIENTO EN LA UNIVERSIDAD, Nómadas (Col), núm. 18, mayo, 2003, pp. 195-202, Universidad Central, Bogotá, Colombia.

[11] Carrasco Gallego, Amalia; Donoso Anes, José Antonio; Duarte-Atoche, Teresa; Hernández Borreguero, José Julián; López Gavira, Rosario; Diseño y validación de un cuestionario que mide la percepción de efectividad del uso de metodologías de participación activa (CEMPA). El caso del Aprendizaje Basado en Proyectos (ABPrj) en la docencia de la contabilidad; INNOVAR. Revista de Ciencias Administrativas y Sociales, vol. 25, núm. 58, octubre diciembre, 2015, pp. 143-158; Universidad Nacional de Colombia, Bogotá, Colombia.

[12] Mejía Murillo, Carmen, Manual de Procesos de Investigación Formativa, Universidad Herminio Valdizán, Perú, 2016.

[13] Pinto Santos, Alba, Cortés Peña, Omar, ¿Qué Piensan los Estudiantes Universitarios Frente a la Formacion Investigativa?, REDU. Revista de Docencia Universitaria, 2017, 15(2), 57-75.

[14] Baluarte Araya, César., Vidal Duarte, Elizabeth, Castro Gutierrez, Eveling, Validación de las Habilidades Blandas en los cursos de la Currícula de la Escuela Profesional de Ingeniería de Sistemas-UNSA, 16th LACCEI International Multi-Conference for Engineering, Education, and Technology: “Innovation in Education and Inclusion”, 19-21 July 2018, Lima, Perú. Digital Object Identifier (DOI):

http://dx.doi.org/10.18687/LACCEI2018.1.1.97 ISBN: 978-0-9993443-1-6 ISSN: 2414-6390.

[15] Baluarte, C., Vidal, E., Delgado, L., Castro, E.; Integrando Habilidades Blandas: Redacción, Comunicación y Ética en la Currícula de la Escuela Profesional de Ingeniería de Sistemas – UNSA; 15th LACCEI International Multi-Conference for Engineering, Education, and Technology: "Global Partnerships for Development and Engineering Education", 19-21 July 2017, Boca Raton Fl, United States. Digital Object Identifier (DOI): http://dx.doi.org/10.18687/LACCEI2017.1.1.141 ISBN: 978-0-9993443-0-9 ISSN: 2414-6390.

[16] Castro, E., Vidal, E., Baluarte, C., Integrando la Comprensión de la Responsabilidad Ética y Profesional en una Carrera de Ingeniería: Experiencia y Lecciones Aprendidas, 14th LACCEI International Multi-Conference for Engineering, Education, and Technology: "Engineering Innovations for Global Sustainability", 20-22 July 2016, San José, Costa Rica, RP139.

[17] Universidad Politécnica de Madrid. (2008). Aprendizaje orientado a proyectos. Recuperado de: http://innovacioneducativa.upm.es/guias/AP_PROYECTOS.pdf

[18] Dirección de Desarrollo Curricular y Docente, Universidad de La Frontera, Manual de orientaciones: Estrategias Metodológicas de Enseñanza y Evaluación de Resultados de Aprendizaje, Chile, 2018.

[19] Marco A. Villalobos-Abarca, Marco, Herrera-Acuña, Raúl A. Ramírez, Ibar G. Cruz, Ximena C., Aprendizaje Basado en Proyectos Reales Aplicado a la Formación del Ingeniero de Software, Formación Universitaria, Vol. 11(3), 97-112 (2018) http://dx.doi.org/10.4067/S0718-50062018000300097, UTA, Chile.

[20] Rodríguez Andara, Alejandro, Río Belver, Rosa, Larrañaga Lesaka, José María; Aprendizaje Basada en Proyecto (PBL), descripción de una experiencia desarrollada en aula universitaria y sugerencias para optimizar resultados, Universidad del País Vasco, España.

[21] Octaedro, Rúbricas para la evaluación de competencias, España, 2013.

[22] Cano, Elena, Las rubricas como instrumento de evaluación de competencias en educación superior uso o abuso, Profesorado, VOL. 19, Nº 2 (mayo-agosto 2015), Universidad de Barcelona, España.

[23] Sáiz Manzanares, Maria Consuelo y Bol Arreba, Alfredo; Aprendizaje basado en la evaluación mediante rúbricas en educación superior, Elsevier, Suma Psicologica, SUMA PSICOL. 2014;21(1):28-35, España, 2014.

[24] Quaas Fermandois. Cecilia, Nuevos Enfoques en la Evaluación de los Aprendizajes, Revista Enfoques Educacionales Vol.2 Nº2 1999-2000 Departamento de Educación Facultad de Ciencias Sociales Universidad de Chile.

[25] Ortega Andrade NA, Romero Ramírez MA, Guzmán Saldaña RME, Rubrica para evaluar la elaboración de un Proyecto de Investigación Basado en el Desarrollo de Competencias, Universidad Autónoma del estado de Hidalgo, México. https://www.uaeh.edu.mx/scige/boletin/icsa/n4/e6.html

[26] Bautista-Vallejo, José, Espigares, Manuel, Hernández-Carrera, Rafael, Aprendizaje Basado en Proyectos (ABP) ante el reto de una nueva enseñanza de las, Revista Brasileira de Ensino de Ciencia e Tecnología, Ponta Grossa, v. 10, n. 3, p. 43-60, set./dez. 2017. DOI: 10.3895/rbect.v10n3

[27] Sanmartí, N. y Márquez, C. (2017). Aprendizaje de las ciencias basado en proyectos: del contexto a la acción. Ápice. Revista de Educación Científica, 1(1), 3-16. DOI: https://doi.org/10.17979/arec.2017.1.1.2020.

Heart Rate Monitoring with Smart Wearables using Edge Computing

Stephen Dewanto[1], Michelle Alexandra[2], Nico Surantha[3]
Computer Science Department
BINUS Graduate Program – Master of Computer Science
Bina Nusantara University, Jakarta, Indonesia

Abstract—**Heart is a vital component of every human health. The development of wearables and its sensor enables the possibility of easy-to-use real-time monitoring. The goal of this study is to improve an IoT monitoring system by enabling real-time heart rate monitoring and analysis, also to assess the use of PPG sensors in smart wearables compared to other clinical-tested heart rate sensors. The PPG sensor will be used to record heart rate data of the user physically. The measurements are then sent to the application for pre-processing. The application can then transmit the pre-processed measurements to the cloud server for monitoring or further analysis, i.e. to assess the health of users' heart. The measurement comparison with measurement collected by a BCG sensor is carried out in this paper. While neither are standard for heart rate measurements, the findings of the evaluation show that the PPG sensor achieves quite similar input data and assessment results during awake stages. The Fitbit sensor tested often underestimates, with sometimes delayed or doesn't detect a sudden increase in heart rate during sleeping.**

Keywords—Internet-of-things; heart rate; smart wearables; real-time monitoring

I. Introduction

Cardiovascular health is vital to individual's health and performance. Several researches have studied the relationship between cardiovascular and physical activity [1]-[3]. Several more researches that study people with bad habits such as smoking or drinking with their cardiovascular health and along with other health problems [4]-[5].

It's recommended to get enough exercise to improve cardiovascular health and reduce the risk of getting cardiovascular disease. Author in [6] suggests an increase of physical fitness during childhood can improve cardiovascular health during adolescence. Another research studies the relationship of sleep problems and cardiovascular disease [7]. As a prevention measure, a method or mechanism is needed to be able to track cardiovascular health by monitoring heart rate.

Currently, heart rate (HR) monitoring often involves the use of electrocardiography (ECG) and often used for polysomnography (PSG). PSG is a multi-parametric test for quantifying sleeping quality from multi sensors and is adaptable. PSG sensors typically consist of electroencephalography (EEG), electrooculography (EOG), electromyogram (EMG), and ECG. The use of ECG sensors requires the assistance of an expert and often requires body contact, which can disturb the sleeping process itself.

Newer technology aims to eliminate the disturbance caused by sensors, such as ballistocardiograph (BCG) sensor. In 2019, Surantha et al. studied the use of BCG sensor along with developing an IoT network [8] which consists of microcontrollers and web applications for sleep quality monitoring. The use of BCG sensor introduces the possibility of daily and portable monitoring for sensor data.

Meanwhile, installation of BCG sensor and the IoT system proposed in previous study requires the assistance of an expert. The installation of the sensor requires a static placement and sensor calibration before use. It is also focused to sleep-related activities and problems.

This scientific study has been carried out in response to previous research which still needs studies for practical sensor that could improve and/or replace the use of BCG sensors. The result of this research is expected to provide options to set up individually a real-time monitoring system and to provide reports for medical experts periodically using IoT. The presentation of the result from the extracted health data itself should not be treated as a medical advice from an expert and should only be used as analytical data.

The remainder of the paper is structured as follows. The related works are listed in Section II of this paper. The Section III explains the background material and research methodology. The system designs and the simulation results are explained and evaluated in Section IV. Finally, the conclusion is presented in Section V.

II. Background Materials

Some of the background materials and research methods shown in Fig. 1 are discussed in this section.

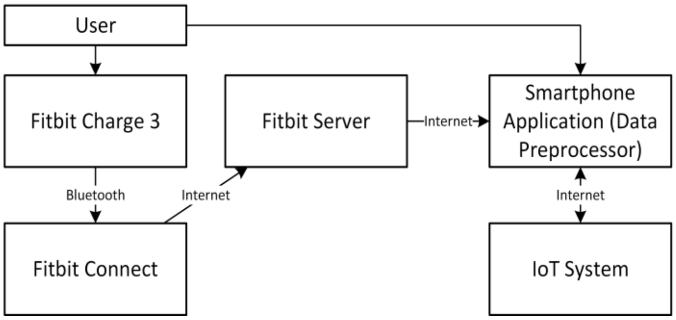


Fig. 1. Background Materials.

A. Fitbit Charge 3

The Fitbit Charge 3 is a fitness tracker that can track activity, nutrition, and sleep by using inputs from user such as weight, drinks, food, or inputs from sensors for measuring heart rate, and location. The sensor is using a PPG signal, where a pulsatile waveform measures relative blood volume changes from blood vessels that is located close to the skin [9]. The waveform is superimposed with a lower frequency one that usually corresponds to respiration and nervous activity.

B. Smartphone Application (Data Preprocessor)

In this study, data preprocessor is implemented by an Android application. The Fitbit Charge 3 is a tracking-only device, and the only way to get sensor data from a Fitbit Charge 3 is by utilizing built-in Fitbit APIs. The Fitbit Charge 3 is designed to be automatically synced to Fitbit servers using Android's Fitbit application, or Fitbit Connect for desktop and microcontrollers. After getting authorization using OAuth 2.0 for obtaining data, the data is then downloaded and preprocessed for the same output from the Data Concentrator. This enables Fitbit to act as an optional input the IoT system.

C. IoT System

Previously, Utomo et al. (2020) developed an IoT system using an ECG sensor as input [10]. The system is developed further [11] for performance optimizations and the use of BCG sensor as input. The BCG sensor uses Murata's SCA11H sensor and all measurements taken is preprocessed first by the Data Concentrator, before being sent to the IoT system. The system contains a cloud server that acts as cloud database and web application that retrieves sensor data and serves sleep report to user and experts.

D. OAuth

The Fitbit-based application requires user authorization before being able to obtain data using OAuth 2.0 protocol. The OAuth 2.0 is one of the popular protocols [12] mostly used for authorizing single sign-on (SSO). Currently, there are 4 OAuth grant types depends on usage and purposes of the authorization. Both client and server who uses OAuth 2.0 protocol must adhere to best practice guidelines.

E. Polling

Polling is a technique that simulates server push using asynchronous requests [13] as to make events and updates delivered quickly to client. Polling works by sending a periodic request to a specific API endpoint and compares the response to check whether a server event has happened or not within a specified interval. Polling works on top of web protocols such as HTTP, so benefits of using HTTP protocols apply. The only downside is the required overhead for HTTP protocol header size compared to WebSocket protocol [14]. There are two types of polling for server-side implementation. Fitbit uses short polling, which returns a response despite no new updates and in turn requires a predefined interval. Meanwhile, a long polling implementation will hold the server from sending a response until a server event happened.

F. Long Short-Term Memory

Recurrent Neural Network (RNN) excels at making prediction or classification through a time-series based input data. This is because there is a feedback of the next iteration from the previous result as input. The negative effects are that RNN requires a longer time for training and some does not work at all. Long short-term memory (LSTM) addresses this issue by removing unused gradient from each iteration feedback [15] with the implementation of gates in each LSTM "cell".

III. Related Works

There are several methods that have been researched and developed further to monitor heart rate using portable, compact, and less complex sensors.

Montgomery-Downs et al. (2011) [16] created a study to compare Fitbit devices in terms of sleep efficiency calculations along with PSG and actigraphy, using each built-in sleep calculation algorithms. The Fitbit system relies on PPG and accelerometer sensor to infer sleep or wake and classification of sleep stages. Both Fitbit and actigraphy calculations are overestimated, with Fitbit calculations still has discrepancy even when compared to actigraphy.

Paalasmaa et al. (2012) [17] proposed method uses a force sensor that detects heartbeat intervals and respiration cycle lengths. The data was compared with ECG reference with 99% precision and 88.73% recall. The sensor is connected to a microcontroller where the data can then be analyzed locally. The analysis consists of sleep stage classification, stress reactions, heart rate curve and average heart rate, and restlessness index. The result can be accessed with a web application where it is presented as graphs and other statistical data.

Santos et al. (2016) [18] proposes an IoT system for health gateway integrated with intelligent personal assistant (IPA). The system autonomously collects measurements from the user when required from the IPA. The IPA has a specific set of algorithms for required actions and alarm alert that can be defined. The proposed method can only do specific actions depending on the capability of IPA.

Pandey (2017) [19] proposes a machine learning model for prediction and detection of stress based on heart rate using an IoT system. The proposed method pushes raw data from the sensor directly to an IoT system setup in the cloud. The prediction and detection of stress uses logistic regression and support-vector machine (SVM) with an accuracy of 66% and 68% respectively during test.

Li et al. (2017) [20] proposes a monitoring system based on IoT for heart rate. The interval of the IoT monitoring system is based using predefined risk category, where higher risk requires real-time monitoring and lower risk requires an event trigger. The proposed method uses ECG as the sensor and shows that the network quality demand is higher for real-time high-risk user.

Araujo et al. (2018) [21] proposes ApneaLink, which is a portable PSG sensor that is used to read data while the user is sleeping and processes the raw data obtained into a sleep apnea analysis. The proposed method requires the ApneaLink sensor to be bound to the chest and user's stomach. While the analysis happened real-time, it is done offline and only enables the user to monitor themselves.

Utomo et al. (2019) [10] proposes an IoT platform to analyze the sleep classification or stages of sleep as a measure of sleep quality. The proposed method uses ECG signals that is streamed real-time to the platform for monitoring and further analysis.

The continuation of the study by Utomo et al. (2020) [11] proposes an optimization of the platform, along with the use of real-time BCG sensor. The proposed method uses heart rate variability and beat-to-beat data from the BCG sensor. The result of this study introduces a less complex system for real-time monitoring, especially for monitoring heart rate.

Challenged for the ease-of-use to the users and real-time monitoring and analysis, smart wearables heart rate sensor is considered. Consequently, modern smart wearables only require installation of a smartphone application. Most smart wearables heart rate sensor uses PPG sensor technology that can assess heart rate in real-time, but most also requires Bluetooth connection for data syncing.

As it can be seen from previous research and studies, there is a need for easier to setup real-time heart rate monitoring system for regular use. There is also a need for improvement on existing IoT systems to be able to retrieve, store, and process data reliably without hogging too much resources.

Most smart wear sensor technologies have been explored for the possibility of portable monitoring, with ease and comfort in mind along with recurring use that doesn't negatively affect the user. Because of technological advancements of built-in smart wear sensors and increasing level of comfort, it can be considered as an alternative. Thus, the real-time monitoring system using smart wearables sensor is proposed in this research.

IV. Proposed Architecture

The proposed architecture focuses on the use of Fitbit Charge 3 as an alternative to BCG sensor that introduces easy to use HR monitoring system at home, which can then be integrated to IoT system. This study also focuses on using edge computation technology to process data offline. After processing, the data can be sent to the server to be directly used, or further processing. Utomo et al. has developed the system that can monitor sleep quality [11] using real time HR data measured from BCG sensors. This research implements a simpler system with some modifications to allow real-time HR monitoring. This research also compares the HR data obtained between Fitbit Charge 3 and SCA11H BCG sensor.

A. Data Acquisition

For this study, the sensor used is Fitbit Charge 3 PPG sensor, a smart wearable in the form of a wristwatch with Fitbit OS firmware version 28.20001.63.5. The watch will be connected using Bluetooth to a Fitbit account via Fitbit application version 3.14 on Android version 9. After setting up Fitbit Charge 3 connection, the Fitbit application will be synced to Fitbit server. There is an All-Day Sync option in the Fitbit application settings, which when turned on will periodically sync every 15 - 20 minutes.

The challenges while doing the research, is that it is noticeable the All-Day Sync option isn't working well. The workaround is by using Bluetooth-enabled microcontroller with the PC version Fitbit application installed, since the Fitbit smartphone application doesn't seem to work in the background. By linking the same account to both application, Fitbit can now sync seamlessly using either microcontroller or smartphone.

The second challenge is that there will always be a delay about 20 minutes for syncing Fitbit by design. The solution in this proposed design is to delay the data stream for around 20 minutes. By doing that, it will also enable you to use more advanced preprocessing algorithms since there's future data, relatively from 20 minutes ago, of the sensor. This is required for simulating beat-to-beat data given a valid heart rate for a given period.

Lastly, the heart rate data from Fitbit Charge 3 is measured in beat per minutes, while the current IoT system requires beat-to-beat measurement data in milliseconds, also for comparison with SCA11H BCG sensor. The solution is by simulating beat-to-beat data based on future heartbeat per minutes, using long short-term memory (LSTM) recurrent neural network (RNN).

Fig. 2 shows the proposed architecture, and Table I shows available sensor data from Fitbit.

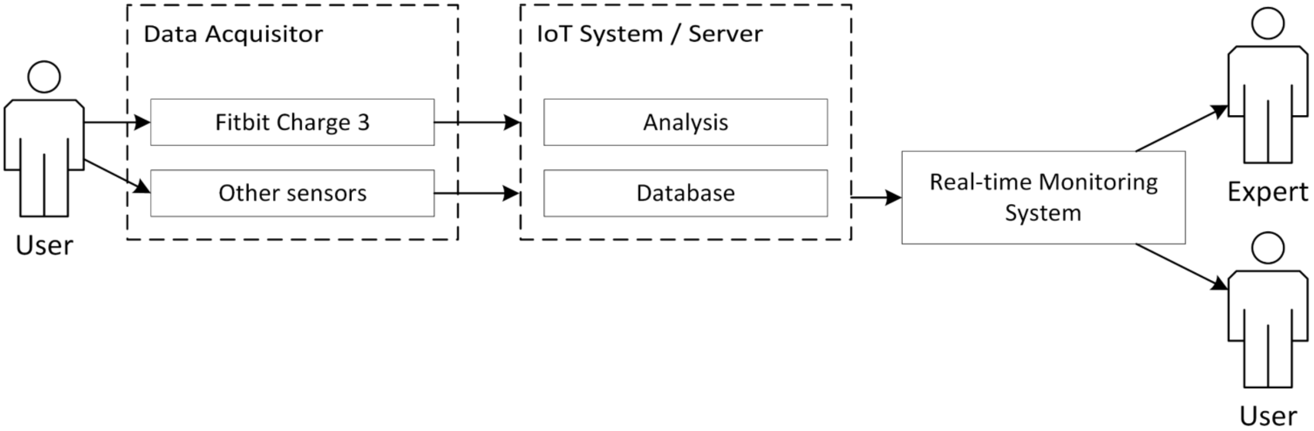


Fig. 2. Proposed Architecture.

TABLE I. MEASUREMENTS FROM FITBIT CHARGE 3

Measure	Data Type	Unit	Description
HR	Integer	1/min	Heart Rate
steps	Integer	-	Number of steps taken
calories_burned	Float	kal	Amount of calories burned
sleep_time	Integer	s	Total sleep time

B. Data Preprocessor

Data Preprocessor is implemented as an Android application. The application uses implicit grant using OAuth 2.0 to Fitbit server, and will start polling for Fitbit sensor data updates. The polling implements a short polling system with an interval of 1 minute to request intra-day heart rate data with interval of 1 second.

Fitbit WebAPIs sends a response containing the specified dataset in a minified JSON format, and subsequent interval with the same heart rate is removed. The Data Preprocessor applies the following algorithm before transmitting data to cloud:

- B2B simulation: Due to the required input of the IoT system and the unavailability of B2B data from Fitbit Charge 3, it is required to predict HRV based on future heart rate per minutes and infer B2B time from predicted HRV. It is done using future time-series HR data with RNN. The input feature is future time-series HR data for a whole minute, passed into an LSTM layer with 8-time steps with relu as an activation function. The output of the LSTM layer is directed to a single dense layer using sigmoid activation function. As the heart rate between person differs, the training and testing should be done separately. Adam is used as an optimizer and the model is evaluated using mean absolute error (MAE) as the loss function.
- Data normalization: The current implementation of IoT system [11] requires B2B and HRV data and concatenates every 3 seconds of data for optimization requirements.

C. IoT System

Data that have been preprocessed by Data Preprocessor will be sent directly to the current implementation of IoT hosted on cloud platform using RESTful APIs. The data is uploaded through a REST interface with the following specification:

- HTTP Method: PUT
- Request body:
    - User ID: string, Fitbit's account ID
    - Timestamp: string, JSON format
    - HR: integer
    - HRV: integer, predicted values through RNN.

Users data can be monitored in real-time with a data delay of around 20 minutes because of Fitbit API system implementation. This study uses a similar IoT system that has been developed [11] with the following hardware and software used:

- 2.4 GHz or 5GHz Wireless LAN 802.11ac for Data Preprocessor, requiring a stable uplink internet connection for data retrieval from Fitbit server and data upload through REST API to the specified IoT system endpoint.
- Fitbit Charge 3, using firmware version 28.20001.63.5, with Android application version 3.14 on Android 9.
- Simple IoT system, hosted in DigitalOcean, with data center located in Singapore. Hardware specifications are: 2GB of RAM, 1 shared CPU unit (vCPU).

V. RESULTS

In this study, an examination was done on 2 healthy adults within age range of 20-25, 1 male and 1 female. The examination will be divided into resting phase and sleeping phase. For the resting phase, examinees are required to sits down and do light or no physical activity for the duration of the examination. For the sleeping phase, examinees are required to sleep normally while wearing BCG sensor and Fitbit Charge 3, with the required monitoring application installed on their smartphone.

Before examination, data are collected from both adults on a 30 minutes rest. The data are then used for training the RNN model to predict HRV values and B2B simulation. The pre-trained model is compiled as a smartphone application which then installed on examinees' smartphone. For the examination, both adults are examined for 30 minutes of rest once a day for a total of 2 days, and 1 night of sleep.

All sensor data are collected using data concentrator for BCG sensor and Fitbit WebAPI for Fitbit sensor. Both data is then compared against each other by aligning the timestamp generated from the sensor. The data is saved as a comma-separated values (CSV) with filename formatted as log-[user]-[rest or sleep]-[n, if any].

A. Integration with IoT System

Integration with existing IoT system can supply another source of real-time data. After authorization through Fitbit server for data access to the examinee, the application will send a HTTP POST request to IoT server REST interface for Fitbit's user ID association. Afterwards, in this study, the only measurement used is the beat per minutes of examinee's heart rate.

After uploading the data to the IoT system, a real-time monitoring is possible. Fig. 3 demonstrates an implementation of IoT system with simple real-time chart for monitoring, where horizontal axis denotes time elapsed in seconds since IoT system started. The system is then tested for 5 hours of real-time monitoring.

The measurement taken for the monitoring dashboard is latency and packet loss. The measurement taken for the Data Preprocessor is packet loss. Both measurements are largely dependent to connectivity.

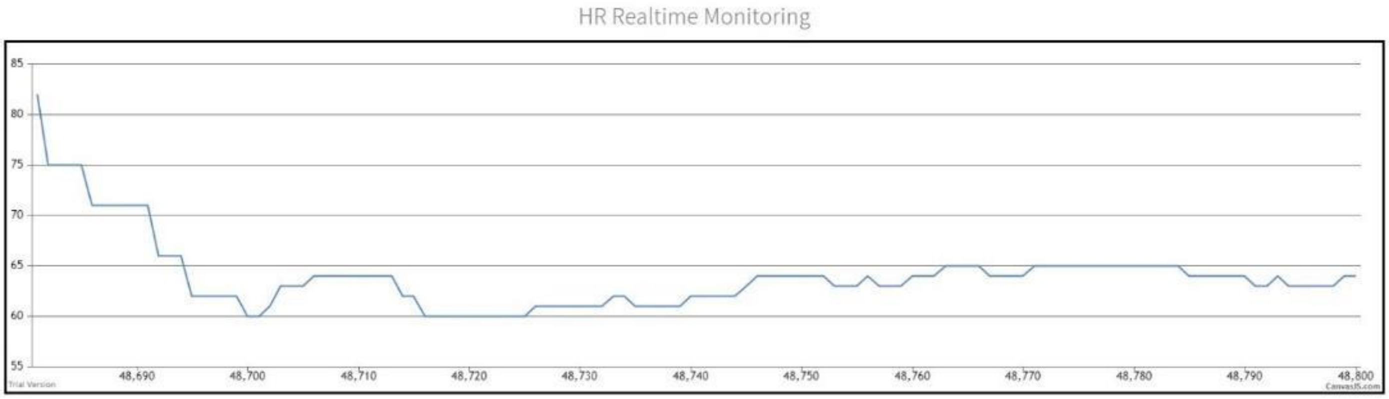


Fig. 3. Real-time HR monitoring Dashboard.

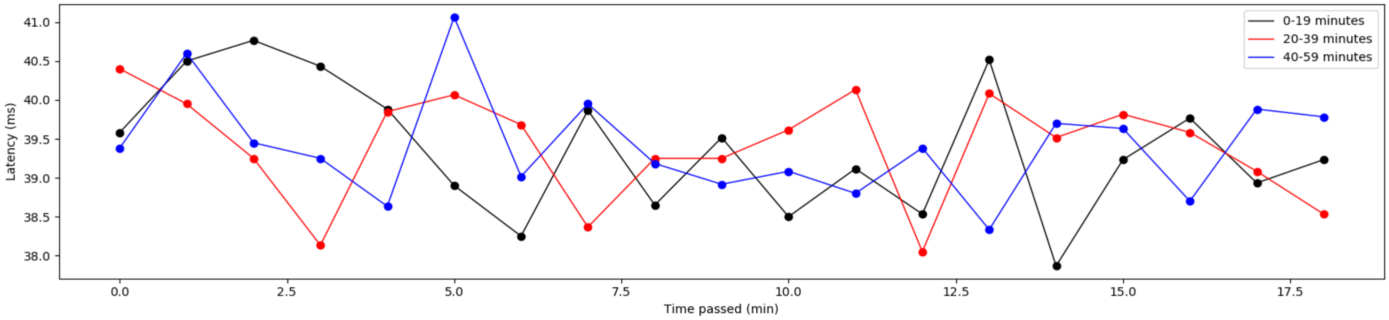


Fig. 4. Monitoring Dashboard Latency.

During the test, there are 0 packet loss through a stable wireless LAN internet connection for both Monitoring Dashboard and the Data Preprocessor, which means all API responses are 200 OK. Meanwhile, the dashboard latency which uses long-polling is around 35-45ms. Fig. 4 shows a detailed latency throughout a random hour of the test for every second in 1-minute average.

The measurements for the server-side IoT system is Disk I/O for storing and retrieving measurements through API, and bandwidth for inbound and outbound traffic. This study implemented a simple IoT system which has little to none impact to memory footprint and CPU usage.

During the test, the IoT system is only running the required process for running the server. Fig. 5 shows server CPU usage, Fig. 6 shows server memory usage, Fig. 7 shows server Disk I/O usage, Fig. 8 shows server bandwidth and Fig. 9 shows average server load. All figures are monitored throughout 6 hours, two ±2.5 hours of test between 1 hour of idle system.

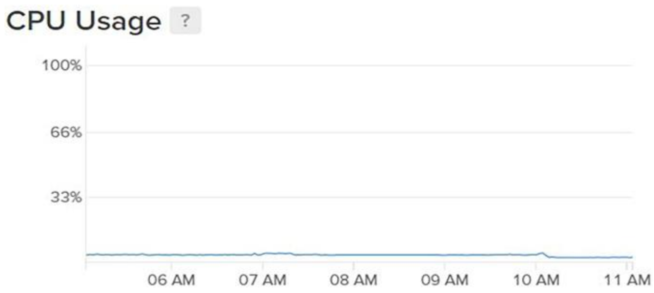


Fig. 5. CPU usage.

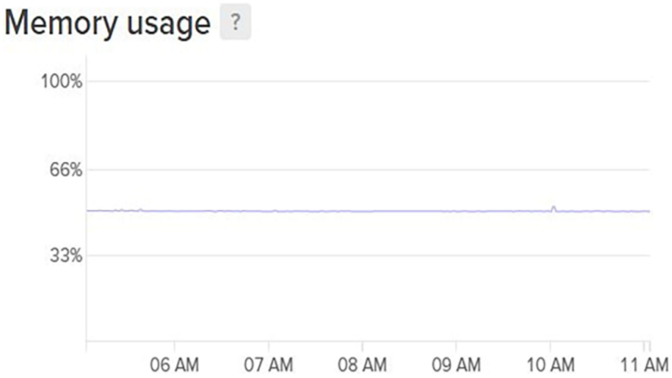


Fig. 6. Memory usage.

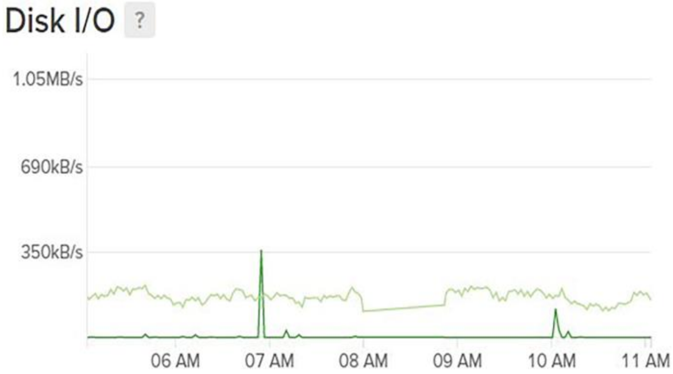


Fig. 7. Disk I/O usage.

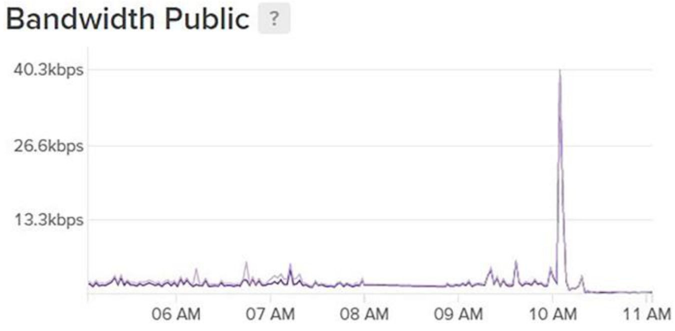


Fig. 8. Bandwidth usage.

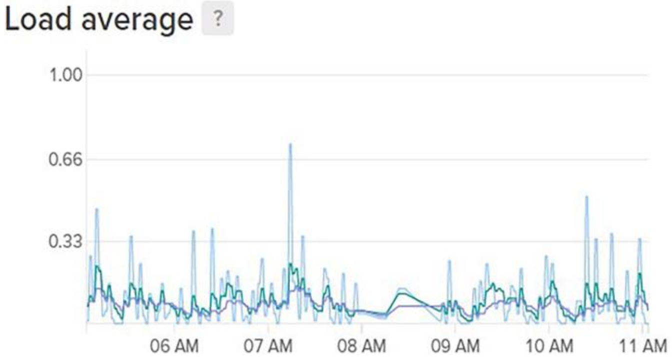


Fig. 9. Average Server Load.

B. Sensor Accuracy

For each examinee, the heart rate measured from Fitbit is compared to the measurements using BCG in beat per minute (bpm). The calculation of error was done using mean square error (MSE) and rooted MSE (RMSE), which tends to give more weight to value that is further from the expected value. Both equations are:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2$$

$$RMSE = \sqrt{MSE}$$

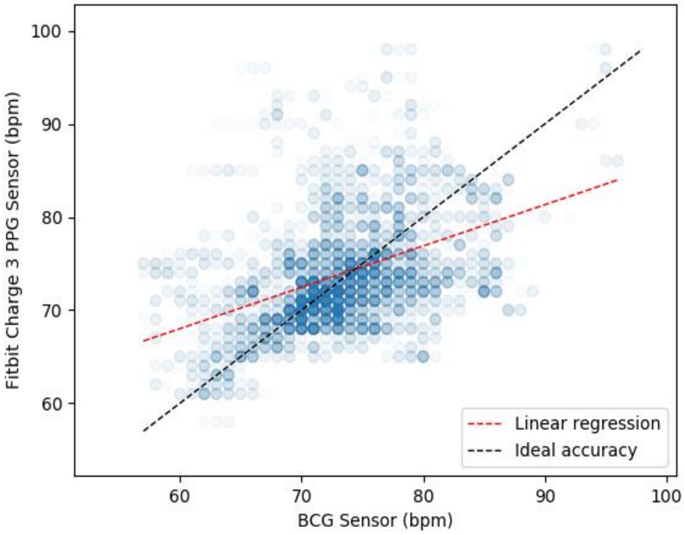


Fig. 10. HR Comparison of Resting Phase.

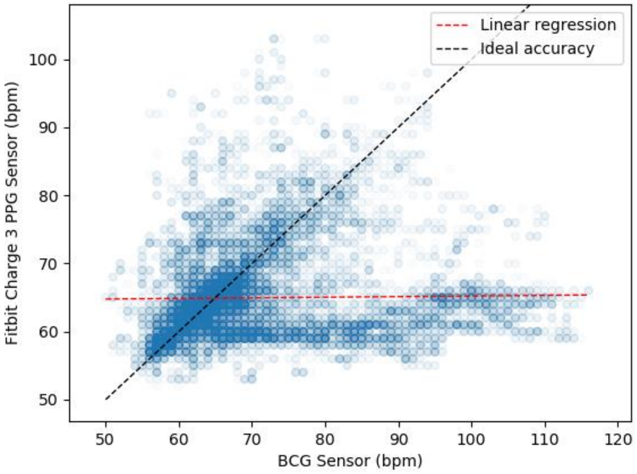


Fig. 11. HR Comparison of Sleeping Phase.

For visualization, the heart rate data is compared within a Cartesian plot, where X-axis represents BCG measurement and Y-axis represents Fitbit measurements. Each point represents a point of time of heart rate measured with BCG sensors and Fitbit with 98% transparency. The ideal accuracy is represented with a black line which represents the same measurements from BCG sensor and Fitbit. The data is separated into resting phase and sleeping phase.

Fig. 10 represents the comparison of BCG and Fitbit Charge 3 during the resting phase and Fig. 11 for sleeping phase.

The MSE of the resting phase for both examinees are measured at 38.58222, when rooted measured at 6.211. The Fitbit tends to overestimate at a maximum of 31 bpm, while underestimate at a maximum of 18 bpm during the examination.

The MSE of the sleep phase for both examinees are measured at 206.45762, when rooted measured at 14.369. The Fitbit tends to overestimate at a maximum of 36 bpm, while underestimate at a maximum of 56 bpm during the examination.

During the sleeping phase, Fitbit method of measuring heart rate seems delayed or even doesn't detect a sudden, increased heart rate for a short period. This makes Fitbit Charge 3 unsuitable to detect disorders or problems during sleeping. While this might be an indicator of cardiovascular problems, both examinees admitted they have never had any, neither in person nor hereditary.

C. RNN Model Evaluation

Recurrent neural networks excel at forecasting using a time-series as inputs, with a total of two differing time-series from examinees sleeping phase. The sleeping phase is chosen for the length and duration of examination, containing more data individually compared to resting phase data combined. The feature selected are time-series data on the next 60 seconds to be predicted the HRV measurement.

Each of these data will be trained separately and evaluated using MAE. Before training, data from BCG sensor needs to be normalized, with the following rules:

- Removal of 0 HR / HRV measurements
- Removal of status 0 [11] which means BCG sensor doesn’t detect signal.
- Removal of greater than 3000 B2B measurements [22], since it is considered as a noise.

The MAE calculates the absolute error without weighting the distance between the error and the expected value. The MAE equation is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

From each examinees’ data, the training will use 75% of the dataset, while using the remaining 25% for testing. The evaluation is measured using MAE for both datasets and result in 2-5% loss for both training and validation dataset. Fig. 12 and Fig. 13 shows detailed training and validation loss during the process for each examinee.

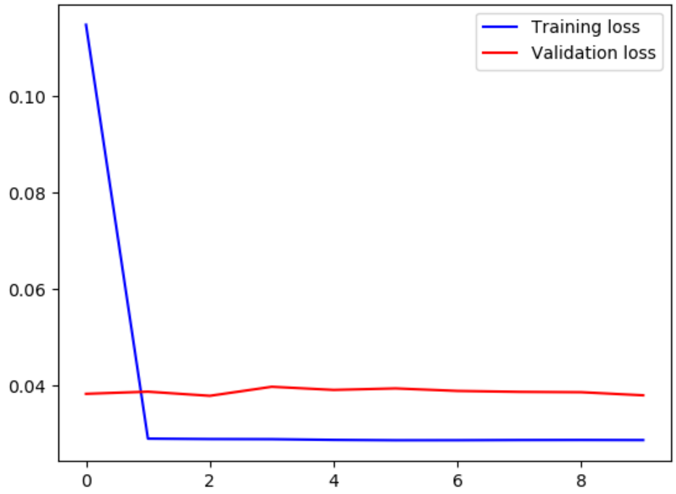


Fig. 12. Training and Validation Loss During RNN Training for Examinee 1.

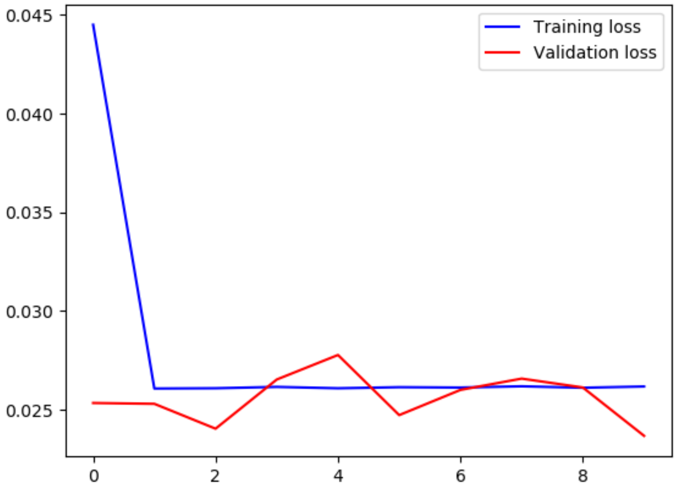


Fig. 13. Training and Validation Loss During RNN Training for Examinee 2.

VI. Conclusions

As observed, built-in heart rate sensors on smart wearables can be an alternative for monitoring heartrate which could be used for other applications such as sleep quality monitoring system. Despite the computing power of most smart wearables, it can be complemented with edge computing. In this study, a smartphone application is used for data preprocessing as an edge computation so the output data can be used directly to the IoT system. The IoT system would store the data for a registered user. Afterwards, an expert or the user can directly monitor user condition through web dashboard in real-time and the ability to view users' historical condition.

Although depending on the problem, the proposed method might prove impractical or inadequate when used as a solution. The proposed method uses Fitbit Charge 3 for its simplicity, security, and robust API documentation for developers. But due to unavailability of raw unprocessed data on Fitbit Charge 3, the Fitbit Charge 3 is not recommended for a more complex application such as sleep quality monitoring which requires HRV values and B2B data. Although this study proposes using RNN to predict HRV values and simulate B2B data using future time-series of measured HR, it is impractical and still requiring further studies.

Meanwhile, there might be other smart wearables with built-in heart rate sensors where it is possible to develop an application directly on the wearables to get and preprocess raw sensor data. While it requires more time to develop such application considering the limited computational power of smart wearables, it is more practical for the users and opens up the possibility of implementation for complex applications.

Currently, our research interest is focused on improving the IoT system by looking for an alternative sensor to get the required input data. The ideal sensor should be one that is practical and easy-to-use, without any discomfort when used regularly for a prolonged period. The objective of this study is also to reduce burden on the IoT system by taking advantage of edge computing for each node of the architecture.

On the other side, it is equally important to analyze the impact of modifications to the performance of IoT system and how much computational power can be distributed to the edge. By doing so, it would allow a balanced usage between edge computational power and improved IoT system efficiency and performance.

Acknowledgment

This work is supported by the Directorate General of Strengthening for Research and Development, Ministry of Research and Technology, Republic of Indonesia as a part of Penelitian Dasar Unggulan Perguruan Tinggi Research Grant to Binus University entitled “Sistem dan Aplikasi Portabel Pendeteksi Gangguan Irama Jantung Menggunakan Sinyal ECG Berbasis Machine Learning dan Cloud Computing” or “Portable ECG-based Heart Arrhythmia Prediction System and Application using Machine Learning and Cloud Computing.”

References

[1] A. K. Gulsvik, D. S. Thelle, S. O. Samuelsen, M. Myrstad, M. Mowé, T. B. Wyller, "Ageing, physical activity and mortality—a 42-year follow-up study," International Journal of Epidemiology, vol. 41(2), Apr 2012, pp. 521-530.

[2] J. N. Ceasar, T. M. Powell-Wiley, M. Andrews, C. Ayers, B. Collins, S. Langerman, et al., "Abstract P395: Examining the Relationship Between Physical Activity Resources and Self-reported Vigorous Physical Activity in a Resource-limited Community: Data From the Washington DC Cardiovascular Health and Needs Assessment," Circulation, vol. 139(Suppl_1), Mar 2019, pp. AP395-AP395.

[3] S. M. Shortreed, A. Peeters, and A. B. Forbes, "Estimating the effect of long-term physical activity on cardiovascular disease and mortality: evidence from the Framingham Heart Study," Heart, 99(9), May 2013, pp. 649-654.

[4] I. S. Ockene, N. H. Miller, "Cigarette smoking, cardiovascular disease, and stroke: a statement for healthcare professionals from the American Heart Association," Circulation, vol. 96(9), Nov 1997, pp. 3243-3247.

[5] E. G. Wilmot, C. L. Edwardson, F. A. Achana, M. J. Davies, T. Gorely, L. J. Gray, et al., "Sedentary time in adults and the association with diabetes, cardiovascular disease and death: systematic review and meta-analysis," Diabetologia, vol. 55, Aug 2012, pp. 2895–2905.

[6] K. F. Janz, J. D. Dawson, and L. T. Mahoney, "Increases in physical fitness during childhood improve cardiovascular health during adolescence: the Muscatine Study," International Journal of Sports Medicine, vol. 23(S1), May 2002, pp. 15-21.

[7] T. D. Bradley, J. S. Floras, "Obstructive sleep apnoea and its cardiovascular consequences," The Lancet, vol. 373(9657), Jan 2009, pp. 82-93.

[8] N. Surantha, G. P. Kusuma, and S. M. Isa, "Internet of things for sleep quality monitoring system: A survey," 2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS), Nov 2016.

[9] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," Physiological Measurement, vol. 28(3), Mar 2007, pp. R1.

[10] O. K. Utomo, N. Surantha, S. M. Isa, and B. Soewito, "Automatic Sleep Stage Classification using Weighted ELM and PSO on Imbalanced Data from Single Lead ECG," Procedia Computer Science, vol. 157, Jan 2019, pp. 321–328.

[11] N. Surantha, C. Adiwiputra, O. K. Utomo, S. M. Isa, and B. Soewito, "IoT System for Sleep Quality Monitoring using Ballistocardiography Sensor", International Journal of Advanced Computer Science and Applications, vol. 11(1), Jan 2020, pp. 200-205.

[12] D. Fett, R. Küsters, and G. Schmitz, "A comprehensive formal security analysis of OAuth 2.0," Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Oct 2016, pp. 1204-1215.

[13] E. Stratmann, J. Ousterhout, and S. Madan, "Integrating long polling with an mvc web framework, 2nd USENIX conference on Web application development," Jun 2011, pp. 113.

[14] F. Y. Jiang, and H. C. Duan, "Application research of WebSocket technology on Web tree component," 2012 International Symposium on Information Technologies in Medicine and Education, vol. 2, Aug 2012, pp. 889-892.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9(8), pp. 1735-1780.

[16] H. E. Montgomery-Downs, S. P. Insana, and J. A. Bond, "A Movement toward a novel activity monitoring device," Sleep and Breathing, vol. 16, Sep 2012, pp. 913–917.

[17] J. Paalasmaa, M. Waris, H. Toivonen, L. Leppakorpi, and M. Partinen, "Unobtrusive online monitoring of sleep at home," Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Aug 2012, pp. 3784–3788.

[18] J. Santos, J. J. Rodrigues, B. M. Silva, J. Casal, K. Saleem, and V. Denisov, "An IoT-based mobile gateway for intelligent personal assistants on mobile health environments," Journal of Network and Computer Applications, vol. 71, Aug 2016, pp. 194-204.

[19] P. S. Pandey, "Machine Learning and IoT for prediction and detection of stress," 17th International Conference on Computational Science and Its Applications (ICCSA), Jul 2017, pp. 1-5.

[20] C. Li, X. Hu, L. Zhang, "The IoT-based heart disease monitoring system for pervasive healthcare service," Procedia Computer Science, vol. 112, Sep 2017, pp. 2328-2334.

[21] I. Araújo, F. Marques, S. André, M. Araújo, S. Marques, R. Ferreira, et al., "Diagnosis of sleep apnea in patients with stable chronic heart failure using a portable sleep test diagnostic device," Sleep and Breathing, vol. 22(3), Jan 2018, pp. 749-755.

[22] S. Nurmi, T. Saaresranta, T. Koivisto, U. Meriheinä, and L. Palva, "Validation of an Accelerometer Based BCG Method for Sleep Analysis," Aalto University publication series SCIENCE + TECHNOLOGY, Jun 2016.

Prediction of Prostate Cancer using Ensemble of Machine Learning Techniques

Oyewo O.A[1], Boyinbode O.K[2]
Department of Computer Science
Federal University of Technology, Akure
Ondo State, Nigeria

***Abstract*—Several diseases are associated with humans; some are synonymous to female and some to male. Example of diseases synonymous to the male gender is Prostate Cancer (PC). Prostate cancer occurs when cells in the prostate gland starts to grow uncontrollably. Statistics shows that prostate cancer is becoming an epidemic among men. Hence, several research works have tried to solve this problem using various methods. Although numerous medical research works are ongoing in the area, the need to introduce technology to battle the epidemic is paramount. Because of this, some researchers have developed several models to help solve issues of prostate cancer in men, but the area is still open to contribution. Recently, some researchers have adopted some well-established Machine Learning (ML) techniques to predict and diagnose the occurrence of prostate cancer, but issues of low prediction accuracy, inability to implement model, low sensitivity; among others still lingers. This paper approached these challenges by developing an ensemble model that combines three (3) ML techniques; Support Vector Machine (SVM), Decision Tree (DT), and Multilayer Perceptron (MP) to predict PC in men. Our developed model was evaluated using sensitivity, specificity and accuracy as performance metrics, and our result showed a prediction accuracy of 99.06%, sensitivity of 98.09% and, specificity of 99.54%, which is a relative improvement on the existing systems.**

Keywords—Prostate cancer; machine learning; support vector; machine; decision tree; multilayer perceptron; diseases

I. Introduction

Cancer is considered one of the most dangerous diseases in the world because it is responsible for around 13% of all deaths in the world[1]. Cancer usually starts as being primary to a specific organ in the body, which later metastasizes to other parts. A common type of cancer is the Prostate Cancer (PC)**.**Prostate cancer is the most rampant and leading cause of cancer death among men in the world, second only to leukaemia [2]. Prostate cancer which is medically referred to as carcinoma of the prostate, begins when cells in the prostate gland starts to grow uncontrollably. Research in [3]explained that prostate cancer begins when healthy cells in the prostate gland change and grow out of proportion, thereby forming a mass called tumour. Recent development in artificial intelligence is now being applied to various fields in medicine and science generally. One of these fields is in the use of Machine Learning techniques to solve issues of prostate cancer. Although, several researchers have tried to predict and diagnose PC in men using several well established ML techniques individually, research in [4],[5], and[6], among others, shows that issues of low prediction accuracy and sensitivity still lingers. This research approached these challenges by combining three (3) well established machine learning techniques (Decision Tree, Support Vector Machine and Multilayer Perceptron), to form an ensemble model that aims to address the recurrent issues associated with the use of single Machine Learning techniques.

The rest of this paper is organized as follows: Section I introduces prostate cancer and justifies the need to carry out this research, Section II reviews related works that have attempted to predict and attend to issues relating to PC, Section III explains the methodology, Section IV presents the results and discussion, Section V concludes

II. Related Works

The prevalence of prostate cancer is increasing by the day. Statistics shows that almost one-third of men over 50 years old will be diagnosed with prostate cancer during their life time[7]. Author in[8], defined prostate cancer as the cancer that occurs in the prostate, a small walnut-shaped gland in men that produces the seminal fluid that nourishes and transports sperm. It is recommended that men have a prostate examination by age 50 [9]. Performing prostate test starts with a Prostate Specific Antigen (PSA) test, and a core biopsy is recommended should the patient have PSA value higher than normal [7]. Biopsy is the gold standard for cancer diagnosis.

Although several works have tried to contribute to Prostate cancer epidemic using various medical approaches, the advent of technology also brought about the development of some computer aided solutions. Example is in [7] where the authors developed a computer aided diagnostic tool that uses image processing techniques for efficient PC diagnosis and prognosis. The authors collected images of prostate gland as shown in Fig. 1, and then separate the images into various portions to diagnose prostate cancer.

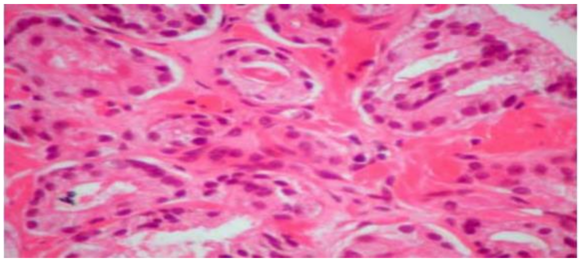


Fig 1. Test Image of Prostate Gland.

The introduction of imaging and machine learning techniques to acquire, process and analyse images from biopsies is of utmost importance[10], because some other diseases imitates prostate cancer. Example is the Benign Prostatic Hyperplasia (BNH), which occurs when the prostate begins to press against the urethra as a result of growth, thereby causing urinary problems[11]. However, the occurrence of prostate cancer is common among men aged 50 and above.

It is essential to trust prediction and diagnosis made using artificial intelligence[12]. Therefore, accuracy of ML predictions is very important. Research in [11] proposed the use of Artificial Neural Network to detect early signs of prostate cancer, but the model could not record perfect accuracy. Author in [13], also applied artificial neural networks (ANN) with back propagation to predict prostate cancer recurrence in patients, but the evaluation could not achieve optimal accuracy. Research in[14]also developed a model using Fisher Linear Classifier to predict recurrence of prostate cancer in men, but their model achieved an accuracy of 93%.Zhao *et al.*,[15]proposed a Penalized Logistic Regression Technique based on top-scoring pair (TSP) as a classification model to predict prostate cancer, but perfect accuracy was also not recorded. Authors in [16]proposed a prostate cancer predictive model using Decision Tree Algorithm. The research established Decision Tree as a useful data mining algorithm for predicting prostate cancer, but the model is not reliable due to low accuracy. Ge*et al.*,[17], proposed a prostate cancer predictive model using Logistic Regression and Artificial Neural Network, but the individual accuracy of the algorithms stood at 84.02% and 85.09% respectively. Takeuchi*et al.*,[18]proposed a prostate cancer prediction system on prostate biopsy using Multilayer Artificial Neural Network (ANN), but the system was able to predict with an accuracy of 71.6%, but this can be associated with the insufficient amount of dataset used for the research. In order to combat the recurrent issue of accuracy, our research proposes an ensemble model that combines three ML techniques. The method and functionality of our model is discussed in the next section.

III. Methodology

The architecture of the model is shown in Fig. 2. The architecture shows the components of the developed model. The functionalities of each component are explained in details below:

A. *Datasets (Prostate and Non-Prostate Cancer)*

The dataset used in the research is obtainable fromhttp://github.com/selva86/datasets/masters/prostate.csv. The obtained data contains about one thousand, nine hundred and forty (1,940) study patients which make up the instances of the data. Each of the instances consist of 10 attributes including class label indicating that an instance is either a Benign (0) or Prostate cancer sample (1). The attribute values are all numeric, Table 1 shows the description of the data attributes.

TABLE I. Description of Data Attributes

S/N	Data Attributes	Description
1	Icavol	Log of the Cancer volume
2	Iweight	Log of the prostate weight
3	Age	Age of the patient
4	Ibph	Log of the Benign prostatic Hyperplasia amount
5	Svi	Seminal Vesicle invasion
6	Icp	Log of the Capsular Penetration
7	Gleason	Gleason Score
8	Pgg45	Percentage Gleason score 4 or 5
9	Ipsa	Log of Prostate Specific Atigen

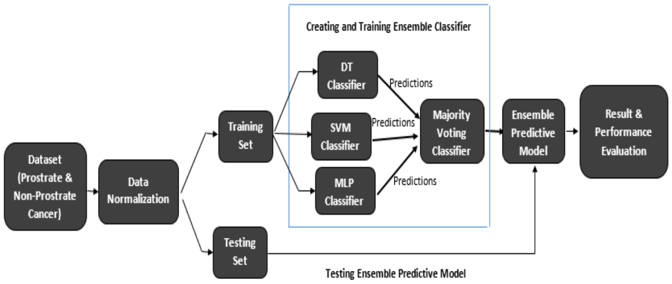

Creating and Training Ensemble Classifier
Dataset (Prostrate & Non-Prostrate Cancer)
Data Normalization
Training Set
Testing Set
DT Classifier
SVM Classifier
MLP Classifier
Predictions
Predictions
Predictions
Majority Voting Classifier
Ensemble Predictive Model
Result & Performance Evaluation
Testing Ensemble Predictive Model

Fig 2. Architecture of Model.

B. *Data Normalization*

The obtained data was normalized using Z -score normalization in order to make training less sensitive to the scale of features.

Z score will convert the data into [0,1] distribution using equation (1)

$$x'_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

Where x'_i is the data value, x_i is the data value to be normalized, μ represents the mean of data values in the feature category

C. *Data Training and Testing*

The normalized data was divided into training and testing set using a 67% - 33% split ratio as shown in Table 2.The training set was used to train the classifiers, the testing set was used to evaluate predictive models.

TABLE II. Prostate Cancer Distribution

Class labels	Training set	Testing set
Non- Prostate	908	432
Prostate	391	209
Total	**1299**	**641**

The classification algorithm used in this research for predicting the presence of prostate cancer is an ensemble of three (3) classifiers: Support Vector Machine (SVM), Decision Tree (DT), and Multilayer Perceptron. The ensemble algorithm predicts the presence of the three classifiers (SVM, DT and MP) predictions $P_1, P_2 and P_3$ respectively, to make final prediction P_f as follows:

Given training set of prostate cancer data is given as:

$$D = \{(x_i, y_i)\}_i^n = 1 \tag{2}$$

Where $D$is the training set of prostate data, x_i is an input for the i -th prostate data described by set of attributes $a_1 a_2 a_3 a_q, y_i \in \{0,1\}$ is its corresponding class label indicating whether the sample is a benign sample ($y_i = 0$) or a prostate cancer ($y_i = 1$), and n represents the total number of data samples.

The first classifier C_1 which is Linear SVM make prediction P_1 as either ($y_i = 0$) or ($y_i = 1$), by creating decision boundaries (hyperplanes) that linearly separates the two classes using equation (3)

$$(w. x_i) + b = 0 \tag{3}$$

Such that

$$\text{Class } (x_i) = \begin{cases} 0, & w.x_i + b \geq 0 \; if y_i = 0 \\ 1, & w.x_i + b \leq 0 \; if y_i = 1 \end{cases} \tag{4}$$

Where x_i denotes an instance of a prostate cancer sample, w represents the weight vector, b is the bias.

However, associated with each hyperplane is a notion called margin, defined as the distance between the hyperplane ($w. x_i) + b = 0$ and the closest sample x_i which can be determined using equation (5)

$$\frac{|w.x_i + b|}{||w||} \tag{5}$$

The best choice of hyperplane depends on the hyperplane with maximum margin between both classes. This is achieved by minimizing weight vector $||w||$ using equation (6)

$$\min ||w||^2, st. \begin{cases} 0, & w.x_i + b \geq 0 \; if y_i = +1 \\ 1, & w.x_i + b \geq 0 \; if y_i = -1 \end{cases} \forall i \tag{6}$$

Decision Tree

The second classifier C_2 make prediction P_2 by applying C4.5 algorithm, as it starts by selecting an attribute from the given set $a_1, a_2, a_3, \ldots., a_q$ to partition D into subsets ($d_1, d_2, d_3 \ldots., d_j$) using information gain presented in equation (7) and (8).

$$I(D) = -\sum^m P(y_i) \log_2 P(y_i) \tag{7}$$

$$IG(D, a_i) = I(D) - \sum_{v \in values(A)} \frac{|D_v|}{n} I(D_v) \tag{8}$$

Where v is a value in attribute a_i, value (a_i) represents all possible values in a_i, D_v represents instances for which a_i has v , n represents number of instances in $I(D)$ and $I(D_v)$ respectively, $P(y_i)$ represents the probability of class y_i in D, m is the distinct number of class values, and j is the number of outcome of test attribute a_i.

The process is continued over each d_i, where $1 \leq i \leq j$, until all elements in each final subset falls under the same class.

Multilayer Perceptron

The third classifier C_3 makes its prediction P_3 as MLP accepts input vector x_i multiplied by a weight vector w_i, and added to a bias b to produce an output$\hat{y}$ using the following equation:

$$\hat{y} = f(\sum_{i=1}^n w_i x_i + b) \tag{10}$$

where n is the number of input-output pairs, and f is a non-linear activation function presented in equation (11).

$$f = \frac{1}{1 + e^{-x_1}} \tag{11}$$

To determine the prediction error of MLP, the Mean Square Error (MSE) function is applied as follows:

$$E(\hat{y}, \mathrm{y}) = \frac{1}{2} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\hat{y} - \mathrm{y})^2 \tag{12}$$

Where $E$is the error function between the predicted class $\hat{y}$ and the target classy

Also, training the MLP by backward propagation involves computing the gradient of the error with respect to w is using chain rule of differentiation as follows:

$$\delta_i \leftarrow dE/w$$

Where δ_i is the gradient descent, w represents weight. Thereafter, w is updated in the direction via the gradient that helps minimize the loss.

1) Majority Voting Classification

This involves combining predictions P$_1$, P$_2$ and P$_3$, of the individual base classifiers C$_1$, C$_2$ and C$_3$ respectively to make a final prediction P_f, by predicting the class label that have been predicted most frequently using equation (13) and (14).

$$C_{r,y} = \begin{cases} 1, if p_i = y_i \\ 0, if p_i \neq y_i \end{cases} \tag{13}$$

$$P_f = \arg max \, i \sum_{i=1}^{q=3} C_{r,y} \tag{14}$$

Where C_r represents the decision of the $r - th$ classifier given class y_i , P_f represents the final prediction by the ensemble, and q is the number of the base classifiers.

2) Evaluation Measures

Our model was evaluated based on three metrics: Sensitivity, Accuracy and specificity. Sensitivity measures the proportion of positives (prostate cancer samples) correctly classified, Accuracy measures the proportion of the total number of correct predictions, specificity measures the proportion of negatives (Benign samples) correctly classified, using:

$$\text{Sensitivity} = \frac{AP}{AP + BN}$$

$$\text{Accuracy} = \frac{AP + AN}{AP + AN + BP + BN}$$

$$\text{Specificity} = \frac{AN}{AN + BP}$$

Where AP = True Positive, AN= True Negative, BP= False Positive, BN = False Negative.

Our ensemble model was implemented using Python 3.7, Spyder python editor via Anaconda Distribution, Excel spreadsheet package, Intel(R) Core(TM) i5-4300U CPU @ 1.90GHz, 2501 Mhz, 2 Core(s), 4 Logical Processor(s).

IV. Results and Discussion

The Confusion Matrix result of the developed prostate cancer prediction model when applied on the test data is shown in Table III. From the study, it is shown that out of 209 actual prostate cancer data and 432 non-prostate cancers from the 641 test data, the model predicted 205 prostate cancer instances correctly, and predicted 4 incorrectly, while also predicting 430 non-prostate cancers correctly with 2 incorrectly. In all, 635 data was correctly classified, while 6 were incorrectly classified.

Table IV shows the total number of correct and incorrect classifications obtained after the developed prostate cancer predictive model had been tested. The total number of incorrect and correct classifications was computed by summing the number of AP, AN, BN and BP for incorrect classifications.

Correct Classification = AP + AN = 635

Incorrect Classification = BN + BP = 6

Table V shows the evaluation of the developed model using Accuracy, Sensitivity and Specificity.

$$\text{Sensitivity} = \frac{AP}{AP+BN} = \frac{205}{205+4} = 0.9809$$

$$\text{Accuracy} = \frac{AP+AN}{AP+AN+BP+BN} = \frac{205+430}{205+430+2+4} = 0.9906$$

$$\text{Specificity} = \frac{AN}{AN+BP} = \frac{430}{430+2} = 0.9954$$

In order to test the efficiency of the ensemble model, the dataset was tested with DT, SVM and MP individually, and the result is presented in Table VI.

The result showed that the developed ensemble model had the highest number of correctly classified instances with 635 instances with number of incorrectly classified instances as zero (6) instances. However, MP also showed to be effective as it correctly classified 626 instances and misclassified 15 instances. From the study, SVM result was not suitable for the purpose of this research work as it correctly classified all non-prostate cancer instances as it predicted all the 432 non-prostate cancer correctly, but wrongly classified all prostate cancer instances with the number of AP recorded as zero (0).

The graphical representation is presented in Fig. 3.

Table VII shows the Accuracy, Sensitivity, and Specificity of the developed ensemble model and the base models. Our ensemble model shows to be the most effective model with an Accuracy of 99.06%, Sensitivity of 98.09%, and Specificity of 99.54% as compared to the result from other models displayed in table. Figure 4 shows graphical representation of evaluation of the proposed ensemble model with the base models.

In order to evaluate the performance of the developed ensemble system, our results were compared with some existing works as shown in Table VIII, in which the developed model shows to be a better model for the prediction of prostate cancer based on its high Accuracy. Fig. 5 shows graphical representation of the comparison.

TABLE III. Confusion Matrix Result of the Developed ensemble Prostate Cancer Detection Model

		Predicted Class	
		Non-Prostate Cancer	**Prostate Cancer**
Actual Class	**Non-Prostate Cancer**	AN 430	AP 2
	Prostate Cancer	BN 4	BP 205

TABLE IV. Number of correct and incorrect classification obtained by the developed prostate cancer prediction model

Number of Test Data	**Correct Classification**	**Incorrect Classification**
641	**635**	**6**

TABLE V. Evaluation of Developed Model using Sensitivity, Accuracy and Specificity

Accuracy	**Sensitivity**	**Specificity**
0.9906	0.9809	0.9954

TABLE VI. Comparison of our Ensemble model with Individual Base Algorithms

Models	**AN**	**AP**	**BN**	**BP**	**Correct Classification**	**Incorrect Classification**
MLP	426	6	9	200	626	15
DT	432	0	45	164	596	45
SVM	432	0	209	0	432	209
Developed Ensemble Model	**430**	2	4	**205**	**635**	6

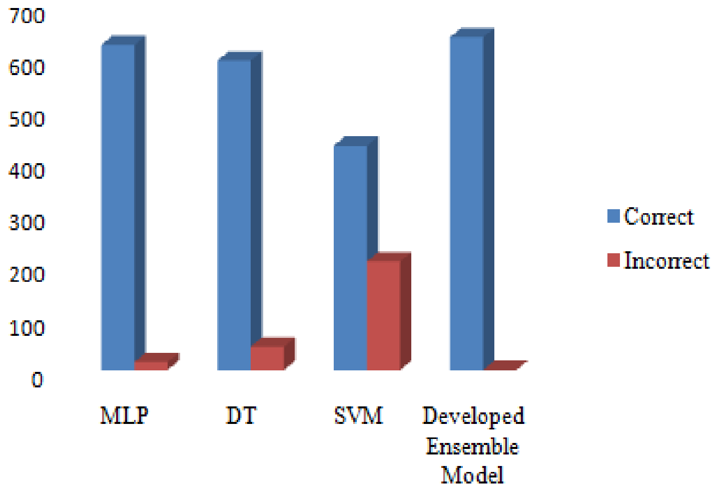


Fig 3. Representation of Correct and Incorrect Classification Ensemble Model with Individual base Models.

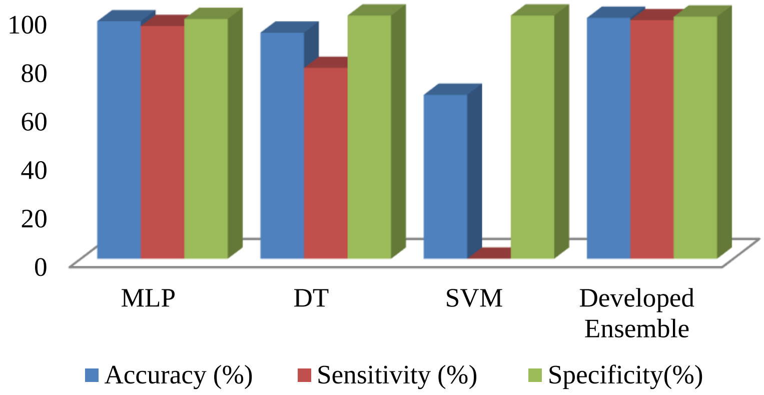


Fig 4. Representation of Evaluation of the Proposed Ensemble Model with Individual base models.

TABLE VII. COMPARISON OF THE DEVELOPED MODEL WITH INDIVIDUAL BASE MODELS

Models	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**
MLP	97.65	95.69	98.61
DT	92.97	78.47	100.00
SVM	67.39	0.00	100.00
Developed Ensemble	99.06	98.09	99.54

TABLE VIII. COMPARISON OF DEVELOPED ENSEMBLE MODEL WITH EXISTING MODELS

Author(s)	**Method /Technique used**	**Accuracy (%)**
Goa and Chen (2015)	Logistic Regression (LR) and ANN	85.09
Xiao et al., (2016)	Random Forest Model	83.10
Takeuchil et al., (2018)	ANN	71.6
Developed Model (2019)	Ensemble of DT, MLP, and SVM	99.06

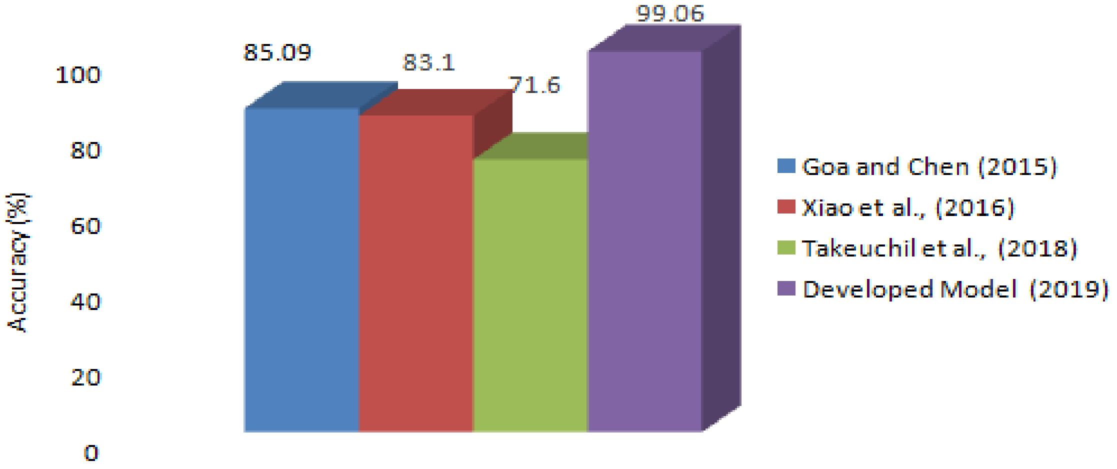


Fig 5. Comparison of Developed Ensemble Model with Existing Systems.

V. Conclusion

The developed model is revealed to be effective in detecting both non-prostate and prostate instances. Using sensitivity, specificity and accuracy as performance metrics, our result has shown a prediction accuracy of 99.06%, sensitivity of 98.09% and, specificity of 99.54%, which is a relative improvement on the existing systems. In other words, we have been able to significantly tackle issues of accuracy and sensitivity in the prediction of prostate cancer in men, using this ensemble model, which shows a relative improvement when compared to the individual base algorithms and some existing models.

References

[1] S. L. Win, Z. Z. Htike, F. Yusof, and I. A. Noorbatcha, “Cancer Recurence Prediction using Machine Learning,” Int. J. Comput. Sci. Inf. Technol., vol. 2, no. 2, pp. 11–20, 2014.

[2] A. Jemal et al., “Cancer statistics 2006,” CA Cancer J. Clin., vol. 56, pp. 106–130, 2006.

[3] T. O. Akinremi, A. Adeniyi, A. Olutunde, A. Oduniyi, and C. N. Ogo, “Need for and relevance of prostate cancer screening in Nigeria,” pp. 6–11, 2014.

[4] E. Alexandratou, V. Atlamazoglou, T. Thireou, and D. Yova, “Evaluation of machine learning techniques for prostate cancer diagnosis and Gleason grading Evaluation of machine learning techniques for prostate cancer diagnosis and Gleason grading Eleni Alexandratou *, Vassilis Atlamazoglou and Trias Thireou George Ag,” Int. J. Comput. Intell. Syst. Biol., vol. 1, no. 3, pp. 298–315, 2010.

[5] T. Takeuchi and K. R. Hospital, “Prediction of prostate cancer by deep learning with multilayer artificial neural,” no. August, 2018.

[6] P. Leydon, F. Sullivan, and F. Jamaluddin, “Machine Learning in Prediction of Prostate Brachytherapy Rectal Dose Classes at Day 30,” in Proceedings of the 17th Irish Machine Vision and Image Processing Conference, 2015, pp. 105–109.

[7] M. Gao, P. Bridgman, and S. Kumar, “Computer Aided Prostate Cancer Diagnosis Using Image Enhancement and JPEG2000,” in Proceedings of SPIE Annual Meeting, 2003, vol. 5, no. August.

[8] S. W. D. Merriel, G. Funston, and W. Hamilton, “Prostate Cancer in Primary Care,” Advances in Therapy, vol. 35, no. 9. Springer Healthcare, pp. 1285–1294, 2018.

[9] D. G. Bostwick and J. . Eble, “Urologic Surgical Pathology, Mosby – year book.” 1997.

[10] O. Saidi, C. Cordon-Cardo, and J. Costa, “Technology insight: will systems pathology replace the pathologist,” Nat. Clin. Pract. Urol., vol. 4, p. 39–45., 2007.

[11] O. E. Ernest, O. Awodele, and O. Ebiesuwa, “Early Detection and Diagnosis of Prostate Cancer using Artificial Intelligence Concept,” Int. J. Comput. Appl., vol. 149, no. 6, pp. 42–46, 2016.

[12] A. O. Ige, A. O. Akingbesote, and A. O. Orogun, “Trust e-Market Environment : : A Review,” Can. open Inf. Sci. Internet Technol. J., vol. 1, no. 1, pp. 1–9, 2019.

[13] L. E. Peterson, “Artificial neural network analysis of DNA microarray-based prostate cancer recurrence,” in Computational Intelligence in Bioinformatics and Computational Biology, 2005, pp. 1–8.

[14] N. Iizuka, “Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection,” Lancet, no. 361, pp. 923–929, 2003.

[15] H. Zhao, S. Qi, and Q. Dong, “Predicting prostate cancer progression with penalized logistic regression model based on co-expressed genes,” in 5th International Conference on BioMedical Engineering and Informatics, 2012.

[16] K. H. Gülkesen, İ. T. Köksal, S. Özdem, and O. Saka, “Prediction of prostate cancer using decision tree algorithm,” Turkish J. Med. Sci., vol. 40, no. February 2009, pp. 681–686, 2010.

[17] P. Ge, F. Gao, and G. Chen, “Predictive models for prostate cancer based on logistic regression and artificial neural network.,” in IEEE International Conference on Mechatronics and Automation (ICMA), 2015.

[18] K. Takeuchi, T., Hattori-Kato, M., Okuno, Y., Iwai, S., & Mikami, “Prediction of prostate cancer by deep learning with multilayer artificial neural network.” 2018.

WoT Communication Protocol Security and Privacy Issues

Sadia Murawat[1], Fahima Tahir[2], Maria Anjum[3], Mudasar Ahmed Soomro[4]
Saima Siraj[5], Zojan Memon[6], Anees Muhammad[7], Khuda Bux[8]
Department of Electrical Engineering, Lahore College for Women University Lahore, Pakistan[1]
Department of Computer Science, Lahore College for Women University, Lahore, Pakistan[2, 3]
Department of Information Technology, Quaid-e-Awam University of Engineering[4, 5]
Science and Technology, NawabShah, Pakistan[4, 5]
Department of Information Technology, University of Sufism and Modern Sciences, Bhitshah, 70140, Pakistan[6, 7]
Riphah Institute of Systems Engineering, Riphah International University, Islamabad, Pakistan[8]

***Abstract*—In this paper, we have proposed a novel approach for the prevention of the Internet of Things (IoT) from fake devices and highlighted privacy issues by using third party Application Program Interface (RestAPI) in Web of Things (WoT). For the ease of life, the usage of IoT devices, sensors, and Radio-Frequency Identifications (RFIDs) increased rapidly. Such as in transport for monitoring vehicles, taxi services, healthcare for patient's health condition monitoring, smart cars, smart grids, and smart homes, etc. Due to this for financial gain attackers are targeting these networks or protocol and adversaries are trying to damage the reputation of the organization or to steal intellectual property. From the last two decades or more, the injection vulnerabilities are more threatening security risks for the web application still exists. The new security challenges occur for the security professional or security researchers in the form of IoT or WoT (Web of Things) communication protocols implementation. These protocol Message Queuing Telemetry Transport (MQTT), Constrained Application Protocol (CoAP), WebSockets, and RestAPI have a different type of security issues. Respectively insertion of fake devices, authentication is not implemented in WebSocket connections, and user's privacy can be leaked with the use of RestAPI without its validation. We have developed a program in Personal Home Pages (PHP) for the detection of new devices in the IoT network. With this, the user's privacy and data will be protected along with some critical security issues of WoT underlying protocols.**

Keywords—Internet of Things; Web of Things; WoT; security issues; privacy issues; protocols MQTT CoAP

I. Introduction

As the IoT devices usage is growing rapidly day by day for the easiness in today's busy life. These devices are used in different areas such as at the motorways for vehicle monitoring, auto fines for law violations, healthcare systems, smart cities, smart grid stations, cab services, and cargo services, etc. These devices or sensors have constrained low computational power, low power storage, and heterogeneous. There is a need fora standard for communication between these devices and secure protocols. Too many organizations are to develop the standard communication protocol for interoperability between heterogeneous IoT devices one of them is the World Wide Web Consortium (W3C) working group has developed WoT Metadata Thing Descriptive (TD). With the existence of different types of interaction protocols, software development languages, and information patterns which creates more complexity with the increasing cost of IoT devices configuration and interoperability [1]. Another end is creating great benefits for too many high-profit gains in the form of smart grid stations, smart homes, and smart cities with implementation of security devices which are estimated more than 10 billion dollars from smart homes only [2].

For the financial gains, user's information theft, and to damage the reputation of organization the attackers are targeting the IoT devices or weakness of WoT communication protocols. There are too many types of communication models in a few of them are using web services standards such as RestfulAPI which is a client-server based model, the second method is used messaging of publishing and subscribe [3]. The WoT provides the facility to use old, current, and newly created tools and methods on the websites for the development of IoT devices with different application usage. With the help of WoT, the interconnection between Things is easier than before. The low-level protocol difficulties can be overcome with the use of web application technologies. Such as Hypertext Transfer Protocol (HTTP) and WebSocket would be used for the IoT devices. And the developer can develop an application that can communicate with IoT devices in the same method as for web services such as RestAPI for payment gateways or mobile applications. With the use of these functionalities, the devices can be accessed from anywhere via the Domain Name Server (DNS). But this method also needs a built-in web server within a constrained network of IoT devices [4]. This WoT architecture does not describe the implementation of the communication method between these Things, but this has simplified the deployment of IoT software applications [5]. Another advantage of this the interfaces and working of Things are explained very well, due to this the information collected from big data cloud and installation of different vendor's devices with their monitoring has been made with low cost and administrative efforts. But this easiness and interoperability between heterogeneous devices, the existing security issues have not been eliminated. At the top of all the web application security issues such as Structured Query Language (SQL) injection, Cross-site Scripting attacks, session hijacking, integrity issues of information, click hijacking, link redirections, and usage of third party Application Program Interface (APIs) with well-known vulnerabilities, etc. Along with these security issues the new problems have occurred with the use of MQTT, WebSocket and CoAP communication protocols between webservers and these devices. Like WebSocketis not supporting the authentication method for communication. If an attacker can get information regarding sensors and webservers are communicating via WebSocket and their parameters. Then he can insert his own devices on that targeted network and capture the useful information or he can perform Denial of Service (DoS) attack by creating too many fake connection requests on that targeted network. Another protocol the MQTT has too many security issues such as authentication, authorization, confidentiality, and integrity. Because this protocol is designed for the low power processing devices to decrease the overhead of processing messages exchange between devices. As the MQTT protocol is used with the applications for process messages incorrectly, some critical security issues can occur like as fake devices insertion, DoS attack, or remote code execution attack [6]. The third one is the CoAP protocol which works at the application layer and similar to HTTP for the compatibility of current web applications. For the efficient performance, enhancement, and low overhead of some critical operations on low power-constrained devices the proxies are used by this protocol. For a secure version of CoAP such as Secure Socket Layer/Transport Layer Security (SSL/TLS) for HTTPS, the Datagram Transport Layer Security (DTLS) protocol is used for communication which is known as CoAPs protocol [7]. But the security of DTLS can be breached as its communication finished at proxies [8]. As the proxies have the functionality of packet holding, replay messages, and manipulation of messages between end-users and servers. Due to this the Man-in-the-Middle (MITM) attack or DDoS attack can be performed by compromising the security of proxy. So in this paper, we will focus on the prevention from the fake devices on the IoT network that is using WoT. The automated program for the detection of fake devices or insertion of any new devices within an IoT network.

The reset of paper is divided as follows: Section 2 will describe the related work done in this area. In Section 3 the WoT architecture model will be discussed. Section 4 we will discuss the security issue of the protocol used under the umbrella of WoT for heterogeneous IoT devices. In Section 5 the proposed solution will be described for the prevention of fake devices. In the last Section 6, this paper will be concluded.

II. Related Work

In the late first decade of the 21st century, the WoT was proposed by the scholar in research, from their onwards too many research work has been done. How to connect these heterogeneous IoT devices with existing web application protocols? With these efforts the WoT architecture and frameworks have been developed, those where changed in the form of the WoT communication methods and working prototypes defined [9]. The usage of RestAPI has been proposed as a good solution for WoT-based communication services [10]. The author [11] have suggested the framework which is known as WoT STORE for allowing upload and discovery of applications used for the W3C-compatible Things. This framework supports two types of applications such as Thing Application (TA) for the facilitation of deployment the Thing action detail by their TD, and Mashup Applications (MA) activating the connections and information extraction from a different type of Things. Similar to the HTTP protocol the CoAP architecture has been proposed by a scholar with proxies for better performance [12]. The method has been developed that is known as Thing discovery, which is used for the division of two sub-issues like as indexing of resources and finding those resources on the search base with keywords or content [13]. To overcome the issue of finding resources the decentralized device discovery method has been added into the CoAP protocol on multicast-based communication [14], which has only functionality of named-based finding. The scholar mainly discussed the security of devices as the main part of that area and considered the layer as secure which is similar to the Transmission Control Protocol/Internet Protocol (TCP/IP) protocol that consists of security systems and techniques for the IoT networks [15]. Another scholar has proposed a solution for confidentiality as the main feature [16] which deals with devices have less processing power should use asymmetric cryptographic algorithms for the authentication process. And this process requires less processing power for this new method of authentication which is based on the function of hashing or OR operations.

For finding out any default protocol settings the new tool has been proposed which are known as *SecKit* [17] and it is a chain of security toolkit process-based. This tool only tries to implement few security policies against the default configuration of the MQTT protocol. To find the vulnerabilities in applications the fuzzing method of testing is used [18] in this process wrong values are inserted into input data fields for targeted applications and after that, these are monitored for their outcomes. The fuzzing is further divided into two main types [19]: mutation-based and generation-based fuzzing. In mutation-based, the testers are using the mutation method on current information samples for creating test cases. In the second type, the test cases are generated from the start for the selected protocols or file formats. The author [20] have used the mutation-based fuzzing for the security testing of applications which are using the MQTT protocol. As per the author to decrease the effort and time for testing the security of the application this method has been selected as compared to the complex process of generation-based fuzzing. This process is focused only on the MQTT protocol application security. The author [21] has developed a Snap4City IoT framework which works on MQTT over TLS. But this protocol has still two main issues first: The MQTT client should work with Transmission Control Protocol (TCP) and the connection should remain active at all times with its broker. Second: the MQTT contain long characters of string names which may not be supported by all IoT device in that network. Too many techniques have been developed already for decreasing the overhead of DTLS headers. One of them is the 6LoWPAN [22] method has been suggested for compressing the Headers to decrease the overhead of DTLS for the CoAPs. Another author said that the DTLS header compressing may create an issue for the security bits of CoAPs. The same as the compressing method has been applied [23]. For the improvement of a handshake between two parties for authentication, digital certificates are used along with DTLS. But this certificate-based authentication considered not a practical approach for the low processing power IoT devices. The researcher [24] has defined that the DTLS is not a good option for the CoAP due to the usage of proxy and at the development time this protocol was not designed for the low processing power devices. As the DTLS uses six messages for the connection which is not ideal for the resource-constrained devices. The authors have suggested the heavy processing should be offloaded for the trusted gateways which are suing the DTLS handshaking based on digital certificates for the IoT devices [25]. The in-line security suites implementation is also known as radio security suites which facilitate the full security stack. The author [26], has added in his in-line security functions for cryptographic processes. As per standard body (IETF) [27], the WebSocket have not a method of authentication and for secure communication. The WebSocket protocols are using the frame-masking technique to prevent proxy cache poisoning attacks, but due to this process security firewalls and sandboxes are unable to detect any malicious data in WebSocket connections [28]. The WebSocket is allowing connection requests to any host and for any TCP port connection request also. By exploiting this functionality, the attacker has to just apply the process of port scanning and network scanning for the organization's local area network for creating a connection request with the targeted user [29]. When the attacker can run subjective JavaScript code inside the internet browser, he is likewise ready to start a WebSocket association with discretionary assistance. After this, the aggressor can use the current WebSocket channel to control the internet browser progressively inside the points of confinement of JavaScript [30]. As per our study a lot of work has been done regarding the WoT underlying protocols security but not for the insertion of the fake device in IoT network. So we have worked in this area with an automated program developed in NodeJS for detection of any new IoT devices in-network and that will generate an alert to a system administrator.

III. WoT Architecture Model

A. *Maintaining the Integrity of the Specifications*

For the communication between heterogeneous IoT devices, the W3C has designed the WoT model [31]. This communication will take place without any consideration of which type of current stack and network protocol is in use. For the detection of different types of IoT network communication interfaces, the WoT TD and metadata patterns have been developed. The devices using a WoT runtime and a WoT scripting API which normalizes the communication between different devices at the same layer can define the network interfaces of current devices or create new interfaces with the help of TD. It moreover supports semantic explanations dependent on connected information [32] supporting incredible hunt and inferencing capacities. The WoT architecture has given three main sections those can be categorized in different configuration and technologies as per requirements of installations at site. The overview of the WoT architecture model is shown in Fig. 1.

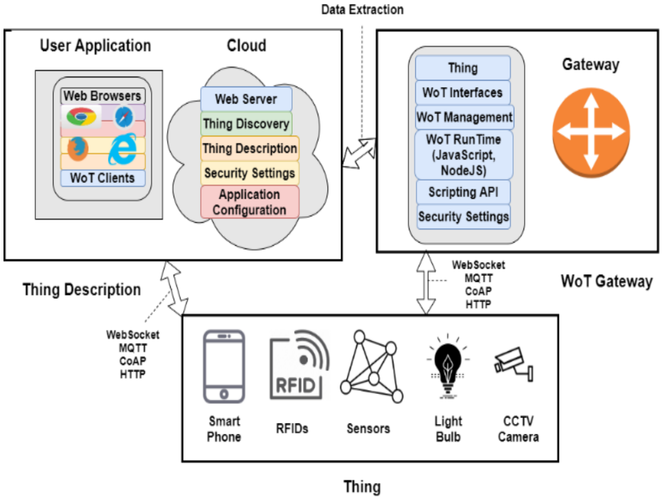


Fig 1. WoT Architecture Model.

B. *Thing*

Application software which may be defined as physical or virtual IoT devices for the communication interface of the network RestAPI. Every Thing is interrelated with the TD [33]. The Thing configuration details, connection types, communication methods, and security settings will be encoded into the TD metadata tag. The WoT Thing works as a server for the networks that only respond to the request but do not start communication itself. Such as a Thing can be a house main gate or electricity controller. That controller may have too many functions for communication which can be operated on that house main gate, for example, to open the door or close it and that may provide communication interfaces with the network to activate these functions.

C. *Thing Description*

An object runs on WoT is known as TD and it activates the communication with their network interfaces. The WoT TD is developed in machine language syntax which gives power to clients for discovery and finds out the functionalities of Things. With these facilities, it is deployed in too many ways as communicate with Things and that will enable communication across the IoT devices. Such as, it may be a web browser or any web application or mobile app on client mobile phone which allows them to activate the communication for given house main gate or electricity controller. The TD supports classic JavaScript Object Notation (JSON) libraries or JavaScript Object Notation for Linked Data (JSON-LD) formats for processing as an information model. The utilization of a JSON-LD processor for handling a TD moreover empowers semantic preparing including the change to Resource Description Framework (RDF) significantly increases, semantic induction and achieving assignments given dependent on ontological terms, which would cause Clients to carry on increasingly independent.

D. WoT Gateway

An object is considered as the client-server setup and it is also known as Servient. It gives single or more WoT Thing interfaces as a server and it will function as a client also for activating communication with other WoT. The WoT gateway provides single or more TDs and related protocol binding processes. This binding process will be started with the TD for specific IoT protocol, like ass HTTP, WebSocket, MQTT, and CoAP. The WoT Runtime and WoT Scripting RestAPI all are hosted at the gateway. For the high-level programming languages (Java Scripting, NodeJS) the WoT scripting RestAPI are deployed for logical operations and this is optional functionality at this layer. Such as it may be a service running on gateway for the smart home and that gives the function of “door locking” services for house main gate, electricity management control, home security alarm with their network interfaces.

IV. WOT PROTOCOLS SECURITY ISSUES

For the easiness of life, the IoT has played a vital role such as for a healthy diet, daily workouts monitoring, patient health monitoring, cab services, vehicle tracking, and much more. By this usefulness, the industry has gain financial benefits which can be considered as the main advantage but at the same time, new security problems have occurred with this fast-growing field. As the high vulnerability security risks are already existing for web applications such as injection type of attacks, session hijacking, data manipulation and more are already defined in Section 1 of this paper. Underlying those security issues new problems have been occurring with the WoT protocols. Which are creating security issues for the user's data leakage or tempering and their privacy?

A. MQTT Protocol

The MQTT protocol operates at the application layer of IoT architecture which depends on applications. This protocol is most widely used for wireless networks and low power processing devices. Its communication is based on the publish-subscribe technique (Fig. 2) and work with the low overhead of packet exchange between communicating parties.

The MQTT protocol is customized for better performance to gather information at the center point and analyze interconnected IoT devices and smartphones for which applications are running on these devices for the datacenter. The smartphone apps are using the MQTT protocol for sending and receiving messages by utilizing an MQTT library. These messages are forwarded by the messaging server of MQTT. After this, the control of delivering messages is transferred to MQTT client and server for the smartphone apps and administration of network for little tasks. As the easy process of implementation and for used most popular applications such as Facebook using it in their chatting app, Amazon for their web services, and many more open source apps or tools are also using MQTT. But at the same time, it has some critical security issues that also should be fixed or prevented for the security of user data and privacy. Few of them are discussed are and we are focused on preventing fake device insertion on IoT network. As major security issues with MQTT, it does not support authentication by default and it can lead to masking any targeted user identification. By doing this an attacker can insert his device and transfer malicious information or capture user’s data. For this, we have proposed an automated program for the detection of any new device insertion in IoT network and that is explained in Section 5.

B. CoAP Protocol

The CoAP is also an application layer protocol that is used in low power processing devices, low storage of battery, and for the IoT networks with limited resources. It is based on a web application protocol model in which the request-response method is used. For the support of current web applications, this protocol has been designed as a copy of the HTTP protocol. And for the best performance, scalability, and decreasing overhead of more processing power-consuming operations on limited resource devices the proxies are being used in this protocol. The CoAPs is known as a secure version of CoAP and the DTLS is used in this version for the TCP layer to encrypt the traffic for two parties. The CoAP overview “Fig. 3” of deployment for the smart city devices or sensors.

As we can see this network has used proxy for the interaction on traditional Internet and that can be compromised by an attacker or prone to cyber-attack. As the communication between two devices or client-server drops at the proxy, it can be captured by an attacker for any malicious intent such as forward fake information, monitor user activates, and spoof user devices and insert his own devices on this network.

C. HTTP Protocol

The HTTP protocol is also application layer protocol and the TCP handshake has been used for the connection establishment between client and server. This protocol works on the request-response method (Fig. 4) for transferring any data.

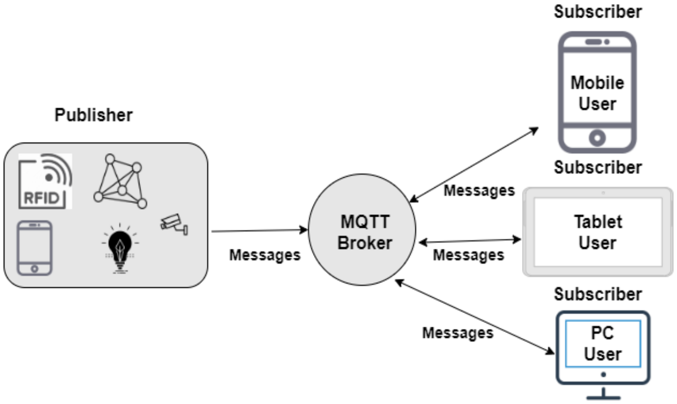


Fig 2. MQTT Protocol.

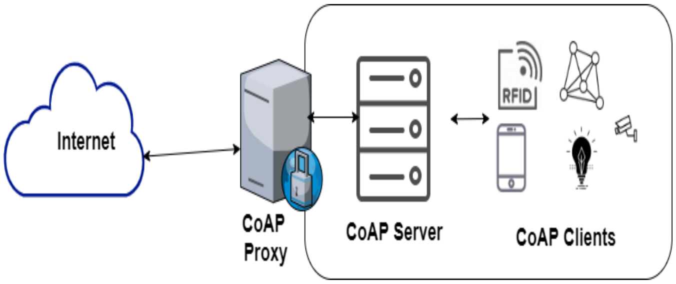


Fig 3. CoAP Protocol Overview.

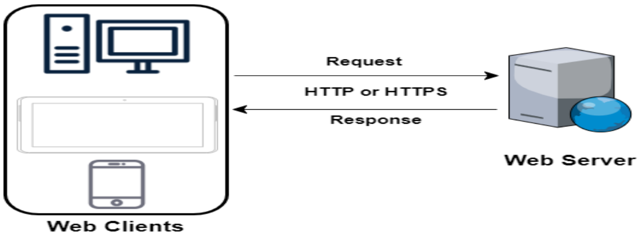


Fig 4. HTTP Protocol Basics.

With the usage of TCP protocol, it can avail all benefits of this protocol such as message delivery authentication, flow control, delivery of messages in proper order, and prevention from congestion [34]. This issue might be probably the greatest obstruction in receiving the web protocols in the usage of WoT for an open IoT environment dependent on open principles. This web application protocol has too many types of security issues such as, click hijacking, injection attacks, third part APIs for known vulnerabilities, and exposing user data if these applications are using an old version of frameworks.

D. WebSocket Protocol

WebSocket is a protocol that communicates in two-way directions for the real-time application on TCP interactions (Fig. 5). As the WebSocket connection has been established between client and server after that they can sync their links to forward information.

At the start of WebSocket development, it was proposed standard along with HTML5 WebSocket API. But now it is developed as a separate entity from HTML5 specs [35]. The WebSocket protocol is a network layer protocol and it is mainly developed for the web browsers and web servers but not limited to these applications it can be utilized in other required services also. The main source for the security of web services is Transport Layer Security (TLS) to scramble traffic and the same policies have been applied for the web browsers as a built-in feature. As we have already mentioned that the WebSocket protocol is different from HTTP so that it can shake the security of web applications. The WebSocket channel allows the attackers for a cache poisoning attack via transparent proxy. To prevent this attack, The WebSocket working group introduced the frame-masking technique. By doing this now firewalls are unable to detect the traffic due to frame-masking and that traffic can be legitimate or malicious. Another security issue with this protocol is it does not provide authentication or scrambling method for communicating parties [36]. Due to this disadvantage, an attacker can exploit it and insert his fake device for monitoring traffic or expose the privacy of users.

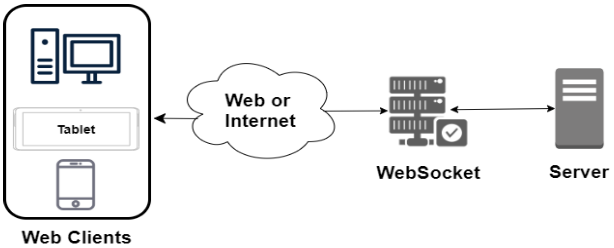


Fig 5. WebSocket Communication Protocol.

V. Proposed Solution and Discussion

There is a big challenge for the intercommunication between heterogeneous IoT devices such as sensors, RFIDs, smartphones, and tracking devices, etc. Currently, too many organizations are to develop a standard method for communication to share the required information between these devices. One of them is W3C have developed WoT architecture by their working group. They have followed the policy of do not reinvent the wheel use already developed protocols for the old and new devices. By doing this the cost of old device replacement will be saved along with administrative efforts. With this benefit of cost-saving and there will be no need to develop new tools and technologies for this rapidly growing IoT networks. At the same time, the too many types of security issues occur some of them are old ones and the new ones also. As authors [37] have used deep earning method for the detection devices on a network. For this they need already data set, images of those devices, and payloads of network transmission. But we are focused on the prevention of fake devices insertion in IoT networks. The program has been developed in PHP for the detection of the new device within a targeted network. With this, the user’s data and their privacy issues will be fixed along with the physical insertion of unwanted devices.

A. A Function for New Device Auto Detection

As the MQTT, WebSocket, and CoAP are more vulnerable to the insertion of fake devices because these protocols are not providing authentication by default. So for the protection of networks from fake devices, we have developed a program for the detection of new devices. The function is given in Fig. 6, will detect new devices as these are inserted. This function will get a *Unique Identification Number* from the connection request of that new device at a targeted network.

After that this will look into databases is that already exists or not. If the record against that device already exists, then operations will be performed normally. If no record exists, then it will save all required information into databases. That new device connection request information will be saved into JSON encoder format regarding connection is established or not and this information will be used for future action against that device.

B. Alert Generation Function

As in the previous function, we have saved connection request information into a database and the same data has been saved into connection file in JSON format. This new device information will be forwarded to the administrator of that targeted network via email (Fig. 7). The administrator will be notified with Short Message Service (SMS) also (we have added dummy email addresses and phone numbers). This process has been applied for as much as quick action against that newly detected device.

With this, we can decrease the damage of data security breaches and user’s privacy. In the current era this too much quick process of notification regarding any activity on the network to administrators.

C. Log Collection and Storage Function

The third function is regarding the storing of new devices connection request activity logs. This function is more important for the checking of any malicious activities and for tracking the footprints. The function will first look for the log file is exists or not. If the log file does not exist, then it will create a new log file for that new device on a system with a *TXT* format (Fig. 8).

Furthermore, this will store all related information to a new connection of that device for future use. This will help a lot to the system administrator for the information regarding how this device has been added to the network and what are the intentions of an attacker. By these all action we can successfully block unwanted devices on our targeted IoT network.

As in this program, the database has been created for the registered devices against their unique number (for example, which may be a serial number, MAC address, or other self-generated unique for these devices). The unique number has been generated from e-tag number, vehicle number, and last year paid tax number. With this, we have blocked fake devices or any tampered e-tag on vehicles by looking into databases with the help of this program.

```
    $this->logNewDevice($uqid); // Save the new device connection is file
            $this->notifySystemAdmin($uqid); // Notifying sytem admin
                        about device connection
                                    }
          echo json_encode(array('success' => 1, 'response' => 'Device
                        connection establised'));
                                } else {
          echo json_encode(array('status' => 0, 'response' => 'Device did
                            not recognised'));
}
  }
 function checkDeviceRegistered($uqid)     {
    $res = $this->db->from('devices')->where(array('uqid' => $uqid))-
                 >get()->row(); // Checking/Fetching in DB
   if (!empty($res))
    return false;
    else
       return true;
```

```
function registerDevice($uqid)
   {
     $this->db->insert('devices',     array('uqid'     =>     $uqid));     //
Adding/Inserting new entery to DB
   }
function deviceConnection()    {
     $uqid = isset($_POST['uqid']) ? $_POST['uqid'] : ''; // Getting the
                 Unique Identification number from request
                                if (!empty($uqid)) {
        if ($this->checkDeviceRegistered($uqid)) { // Checking the device
                already registered in DB              // True
                    //Do something if device already registered

           $this->logDeviceConnection($uqid); // Save the request in the
                                     file
        } else {               // False
                   //If device is not registerd this code will execite
            $this->registerDevice($uqid); // Registering/storing device
                        information in System/DB
```

Fig 6. Auto Detection Function for New Devices.

```
function notifySystemAdmin($uqid)
   {
      $subject = 'Connection Alert';
        $message = 'Device with Unique ID ' . $uqid . ' just got registered
                             with your system';
       sendEmail($subject, $message, 'from@test.com', 'to@test.com'); //
                          Sending Email to Admin
          sendSMS($message, '0xxx2112212'); // Sending SMS to Admin
    }
```

Fig 7. Notification Function for Administrators.

```
function logNewDevice($uqid)
 {
      $file = "new_devices.txt";
            $myfile = fopen($file, "a") or die("Unable to open file!"); //
                 Opening/Getting the file new_devices.txt to log
   $str = "\n\nNew Device with Unique ID " . $uqid . " connected at " .
                              date('Y-m-d H:i:s');
            fwrite($myfile, $str); //Writing/Adding the string/$str to file
   }

   function logDeviceConnection($uqid)
   {
      $file = "devices_log.txt";
      $myfile  =  fopen($file,  "a")  or  die("Unable  to  open  file!");  //
Opening/Getting the file devices_log.txt to log
  $str = "\n\nDevice with Unique ID " . $uqid. " make a connection at " .
                              date('Y-m-d H:i:s');
      fwrite($myfile, $str); // Writing/Adding the string/$str to file
    }
```

Fig 8. Log Gathering and Saving Function.

VI. Conclusion

As the security issues are increasing day by day for the IoT devices and with this, the industry is facing another issue of interoperability between these devices. Which have raised too many questions for the scholars, security professionals, and standard bodies in this area? So that back in 2007 the W3C has proposed framework with the name of WoT and its main architecture has been developed in 2017 and which is still in the development phase. It is a good framework for communication between heterogeneous devices. They have recommended no new protocols but suggested existing and already developed protocols such as MQTT, CoAP, HTTP, and WebSocket. These recommendations were for the current web technologies and web services like RestAPI. As web applications are facing too many critical security issues of injection, session hijack and much more. New security issues have been raised with the use of these IoT protocols. As the MQTT protocol does not provide an authentication method by default. Due to this weakness, an attacker can scan the network for this protocol and if he found it then easily impersonate his device to that targeted network. The CoAP is working similarly as HTTP does and uses proxies for good performance and usability. The secure version of CoAP is known as CoAPs which uses DTLS and these secure connections drop at a proxy. If that proxy server gets compromised then an attacker can do anything on that targeted network like break authentication, forward fake information, and insert his fake device. Another most widely used protocol is WebSocket for the sensors or RFID devices. This protocol also does not provide any authentication method due to this an attacker can easily insert his fake device by just after a successful scan of a targeted network for this protocol. The second issue with this protocol is that its sessions are not closed until the server to client close them. Then this protocol is vulnerable to DoS attack also for too many connection requests. So that we have proposed a novel approach in this paper for the detection of fake devices automatically with the help of the PHP program. The unique numbers are also generated for the detection of tampered devices in case of these are installed at client-side. Our proposed solution has detected fake devices in real-time just by looking into system's databases. This program is generating alerts to administrators via email and SMS for that targeted network. The logs of that new device connection request are also saved at the system for future actions.

ACKNOWLEDGMENT

Authors are thankful to referees for fruitful comments regarding our paper.

REFERENCES

[1] E. Sisinni, A. Saifullah, S. Han, U. Jennehag, and M. Gidlund, "Industrial Internet of Things: Challenges, Opportunities, and Directions", IEEE Transactions on Industrial Informatics vol. 14, no. 11, pp. 4724-34, July 2, 2018.

[2] D. Mary, "http://www.ndpanalytics.com/report-interoperability-and-iot", Last accessed 2020/01/23.

[3] M. McCool, and E. Reshetova, "Distributed Security Risks and Opportunities in the W3C Web of Things", Workshop on Decentralized IoT Security and Standards (DISS), 2018.

[4] D. Guinard, V. Trifa, and E. Wilde, "A Resource-Oriented Architecture for the Web of Things", IEEE Internet of Things (IoT), pp. 1-8, 2010.

[5] Web of Things (WoT) Architecture, "W3C Proposed Recommendation 30 January 2020, https://www.w3.org/TR/wot-architecture/", last accessed 2020/02/04.

[6] K. Kaspersky, and A. Chang, "Remote Code Execution through Intel CPU Bugs", InHack in The Box (HITB), Malaysia Conference, 2008.

[7] E. Rescorla, and N. Modadugu, "Datagram Transport Layer Security Version 1.2", RFC 6347, 2012.

[8] G. Selander, J. Mattsson, F. Palombini, and L. Seitz, "Object Security of Coap (Oscoap)", Internet Engineering Task Force (IETF) Internet-Draft work in progress, 2017.

[9] A. Kamilaris, and M. I. Ali, "Web of Things Platforms" Truly Follow the Web of Things?", IEEE 3rd World Forum on the Internet of Things (WF-IoT), pp. 496-501, 2016.

[10] F. Paganelli, S. Turchi, and D. Giuli, "A Web of Things Framework for Restful Applications and Its Experimentation in a Smart City", IEEE Systems Journal vol. 10, no. 4, pp. 1412-23, 2014.

[11] L. Sciullo, C. Aguzzi, M. Di-Felice, and T. S. Cinotti, "WoT Store: Enabling Things and Applications Discovery for the W3C Web of Things", 16th IEEE Annual Consumer Communications & Networking Conference (CCNC), pp. 1-8, 2019.

[12] L. Mainetti, V. Mighali, and L. Patrono, "A Software Architecture Enabling the Web of Things", IEEE Internet of Things Journal, vol. 2, no. 6, pp. 445-54, 2015.

[13] Y. Zhou, S. De, W. Wang, and K. Moessner, "Search Techniques for the Web of Things: A Taxonomy and Survey", Sensors, vol. 16, no. 5, pp. 600, 2016.

[14] Z. Shelby, K. Hartke, and C. Bormann, "The Constrained Application Protocol (CoAP)", Internet Engineering Task Force (IETF) RFC-7252, 2014.

[15] T. Heer, O. Garcia-Morchon, R. Hummen, S. L. Keoh, S. S. Kumar, and K. Wehrle, "Security Challenges in the IP-based Internet of Things", Wireless Personal Communications, vol. 61, no. 3, pp. 527-42, 2011.

[16] A. Esfahani, G. Mantas, R. Matischek, F. B. Saghezchi, J. Rodriguez, A. Bicaku, S. Maksuti, M. G. Tauber, C. Schmittner, and J. Bastos, "A Lightweight Authentication Mechanism for M2M Communications in Industrial IoT Environment", IEEE Internet of Things Journal, vol. 6, no. 1, pp. 288-96, 2017.

[17] R. Neisse, G. Steri, and G. Baldini, "Enforcement of Security Policy Rules for the Internet of Things", 10th IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), pp. 165-172, 2014.

[18] H. Yang, Y. Zhang, Y. P. Hu, and Q. X. Liu, "IKE Vulnerability Discovery Based on Fuzzing", Security and Communication Networks, vol. 6, no. 7, pp. 889-901, 2013.

[19] M. Sutton, A. Greene, and P. Amini, "Fuzzing: Brute Force Vulnerability Discovery", Pearson Education, 2007.

[20] S. Hernández-Ramos, M. T. Villalba, and R. Lacuesta, "MQTT Security: A Novel Fuzzing Approach", Wireless Communications and Mobile Computing, 2018.

[21] C. Badii, P. Bellini, A. Difino, and P. Nesi, "Smart City IoT Platform Respecting GDPR Privacy and Security Aspects", IEEE Access, vol. 8, pp. 23601-23623, 2020.

[22] S. Raza, D. Trabalza, and T. Voigt, "6LoWPAN Compressed DTLS for CoAP", IEEE 8th International Conference on Distributed Computing in Sensor Systems, pp. 287-289, 2012.

[23] S. Raza, H. Shafagh, K. Hewage, R. Hummen, and T. Voigt, "Lithe: Lightweight Secure CoAP for the Internet of Things", IEEE Sensors Journal, vol. 13, no. 10, pp. 3711-20, 2013.

[24] F. A. Alhaidari, and E. J. Alqahtani, "Securing Communication between Fog Computing and IoT Using Constrained Application Protocol (CoAP): A Survey", Journal of Communications, vol. 15, no. 1, 2020.

[25] R. Hummen, J. H. Ziegeldorf, H. Shafagh, S. Raza, and K. Wehrle, "Towards Viable Certificate-Based Authentication for the Internet of Things", Proceedings of 2nd ACM Workshop on Hot Topics on Wireless Network Security and Privacy, pp. 37-42, 2013.

[26] R. Daidone, G. Dini, and M. Tiloca, "On Experimentally Evaluating the Impact of Security on IEEE 802.15.4 Networks", International Conference on Distributed Computing in Sensor Systems and Workshops (DCOSS), pp. 1-6, 2011.

[27] L. Fette, and A. Melnikov, "The WebSocket Protocol (RFC 6455)", Internet Engineering Task Force, 2011.

[28] M. Shema, S. Shekyan, and V. Toukharian, "Hacking with WebSockets", BlackHat USA, 2012.

[29] S. Shah, "HTML5 Top 10 Threats Stealth Attacks and Silent Exploits", BlackHat Europe, 2012.

[30] M. Schmidt, "HTML5 Web Security 1.0", Compass Security AG, 2011.

[31] K. Kajimoto, M. Kovatsch, and U. Davuluru, "Web of Things (WoT) Architecture", First Public Working Draft, W3C, 2017.

[32] T. Heath, and C. Bizer, "Linked Data: Evolving the Web into a Global Data Space", Morgan & Claypool, Google Scholar Digital Library, 2011.

[33] T. Kamiya, and S. Käbisch, "Web of Things (WoT) Thing Description", W3C, Working Draft, 2017.

[34] N. Naik, and P. Jenkins, "Web protocols and challenges of web latency in the web of things", In Eighth International Conference on Ubiquitous and Future Networks (ICUFN) IEEE, pp. 845-850, 2016.

[35] I. Hickson, "The WebSocket API, W3C Candidate Recommendation, World Wide Web Consortium (W3C)", URL: http://www. w3. org/TR/WebSockets, 2012.

[36] I. Fette, and A. Melnikov, "The WebSocket Protocol", IETF, RFC 6455, 2011.

[37] J. Kotak, and Y. Elovici, "IoT Device Identification Using Deep Learning", arXiv preprint arXiv:2002.11686, 2020.

Smart Energy Control Internet of Things based Agriculture Clustered Scheme for Smart Farming

Sabir Hussain Awan[1]
Department of Electrical Engineering
Iqra National University
Peshawar, Pakistan

Sheeraz Ahmed[2], Atif Ishtiaq[8]
Department of Computer Science
Iqra National University
Peshawar, Pakistan

Zeeshan Najam[3]
Department of Electrical Engineering
MNS Uni of Engg and Technology
Multan, Pakistan

Muhammad Yousaf Ali Khan[4]
Department of Electrical Engineering
FET, Gomal University, D.I. Khan, Pakistan

Asif Nawaz[5]
Faculty of Engineering
Higher College of Technology, Dubai, UAE

Muhammad Fahad[6]
Department of Computer Science
CECOS University of Science and IT, Peshawar, Pakistan

Muhammad Tayyab[7]
Department of Electrical Engineering
Career Dynamic Research Center, Peshawar, Pakistan

***Abstract*—The era of smart farming has already begun, and its consequences for society and environment are expected to be massive. In this situation, Internet of Things (IoT) technologies have become a key route towards new agricultural practices. IoT nodes detect and track physical or environmental conditions and transmit data through multihop routing to their base station. However, these IoT nodes have come up with energy constraints and complex routing processes due to limited capacities. Hence, lead to data transmission failure and delay in the fields of IoT-based farming. Because of these limitations, the IoT nodes distant from the base station are dependent on their cluster heads (CHs), causing additional load on CHs leading to high energy consumption and shortening their lifetime. To address these issues, this research proposes a smart energy control IoT based agriculture clustered scheme to reduce load on CHs by introducing a novel clustering scheme. Simulations are conducted for validation and comparison is made with LEACH protocol in Agriculture and results show that proposed scheme has much lower energy consumption and longer network life as compared to its counterparts.**

Keywords—Agriculture; IoT; network; energy; scheme

I. Introduction

The arrival of the Internet of Things (IoT) is one of the most dynamic and thrilling advances in the information and communication technologies. Though with the passage of time networking technologies have become more universal but they were largely restricted to connecting traditional end-user devices such as mainframes, desktops and laptops and, more recently, smartphones and tablets. Industry analysts estimate that more than eight billion such devices are currently connected to the network and project that this number will grow to over 25 billion by 2020. Some experts are projecting that the IoT could generate revenue of as much as US$ 13 trillion by 2025.Therefore IoT is being applied in every field of life such as smart home, health care, traffic control and smart farming [1].

Smart farming is a management philosophy that seeks to provide the agricultural industry with the infrastructure to use advanced technology for tracking, monitoring, automating and analyzing operations such as big data, the cloud and IoT. Smart farming is also known smart agriculture. Smart agriculture is becoming increasingly important due to rise in world population and increase in food demand. Consequently, it is important to make effective use of natural resources and to increase the use of information and communication technologies to cope with the challenges posed by climate change [2].

The Internet of Things (IoT) is a universal network that enables the monitoring and control of the farm environment through the collection, processing and analysis of the data produced by smart devices to make agriculture smart. IoT smart agriculture helps in decision making cycle which include seed selection, crop select, crop rotation, weeding, watering, harvesting, post- harvesting and pest and disease management [3]. Therefore IoT system can minimize the wastage of crops, efficient use of resources such as water and fertilizers and improve the crop yield and reduce operational expenses [4]. IoT networks for monitoring of farm environment should be of low-cost, making it affordable for farmers and should use low energy for prolong life of the network [5]. There are many sensor nodes in a typical monitoring network, a few sink nodes and a gateway depending on the topology of the network and farm clustering. The sink gathers and uploads data from the sensor nodes onto the server [6]. The sink is always in an active state in most wireless networks, and thus consumes a lot of power [7]. Clustering can help here it is a promising solution which can help alleviate many IoT problems in terms of energy consumption, scalability, usability, etc. due to its similarity to IoT; such as different smart homes can be grouped into different clusters within a standard IoT context, and clustering can also be used by smart devices in a smart home as shown in Fig. 1 [8].

In agriculture for any IoT-based application, the data must be collected by sensing devices and processed by various algorithms and later on the information being processed can be accessed anywhere and at any time via the Internet. The combination of sensing devices is famous as clustering. Clustering provides assistance to efficiently obtain the information with a least number of communications in the network and further transfer information for processing and also supports prolong the lifespan of the network and extend the lifetime of system that is deployed for a particular task.

We also address the clustering of IoT nodes in this research for efficient use of energy and prolong network life to reduce the cost burdon on farmers in the form of system or device replacement. This research presented the IoT network design for the automated collection of soil data from a farm and proposed IoT-based Agriculture scheme, which consumes low energy and has longer life.

In this research we propose a smart energy control Agriculture Scheme based on IoT that uses low energy and has longer network life. The rest of the paper is structured as follows: Section II discusses the Agriculture related work and its current energy-based routing protocols. Section III provides the inspiration for the study. The proposed work design and all measures including clustering process and flow chart are discussed in Section IV. Section V discusses the findings of the simulation and the discussion. The last segment ends the research with conclusion and guidance for the future work.

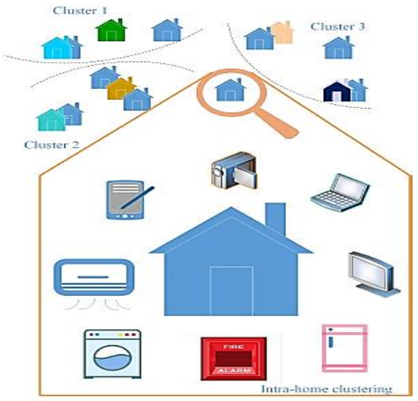


Fig 1. Inter-Home Clustering.

II. Related Work

An e-Agriculture framework proposed by researchers to measures the processes involved with farmers looking for information about agricultural practices to be used and for making decisions throughout the season. It attempted to achieve the information increased amounts that was required to complete and to help growers to accomplish their operations in efficient and effective manner. Initially a framework suggested by decision making theory was that the decision process was a sequence of serial steps that evaluate, identify problems, generate substitute solutions, select and implement them. In addition, these concepts were replaced by a novel and more difficult cycle-based decision-making process [9].

The architecture of the MAC network layer has been developed with periodic data collection, where sensor nodes periodically collect data from fixed locations in the agricultural field [10]. Low-power Wireless communication technology Zigbee was proposed for monitoring agriculture. WSN nodes collect real time data and transmit it to base station using Zigbee. This is a low-cost system where the recorded information is transmitted over an SMS via a GSM network to remote location. The limitation of this setup is its reliance on the GSM network [11]. Authors applied LEACH protocol of wireless sensor network in agriculture to improve the irrigation system. In their research LEACH protocol simulated with three parameters (throughput, end to end delay and total energy consumption of sensor nodes. The draw back within this system that the energy consumption of sensor nodes is high due to which the network life is short [12].

The authors expand the lifespan of the network by splitting the whole area of the network into tiers. In this protocol, sensor nodes of high energy and the nearest distance to the sink are selected as CHs. With the help of an opportunistic multi-path routing system with the goal of minimizing the energy consumed by the selection process of the forwarding nodes to prolong the life of the network [13]. Authors proposed an algorithm based on Distributed Learning Automation (DLA) to boost network life by taking into account various routing constraints such as end-to-end delay and reliability in the selection process of data transmission routes to the base station [14].Researches proposed a diagnostic data collection protocol that considers the clustering and multi-path routing to extend the network's lifetime in IoT network [15] The LoRa Alliance presents the protocol stack for low-power and wide-area Internet of Things (IoT) networking technologies compatible with indoor transmission [16]. Researchers proposed a protocol named as dynamic distributed framework protocol. For the purpose of routing among sensor nodes a mobile agent migration is used for aggregation of data based on energy and trust metric assessment. The drawback of proposed framework supports only small route mobile agent and response time is also low [17].

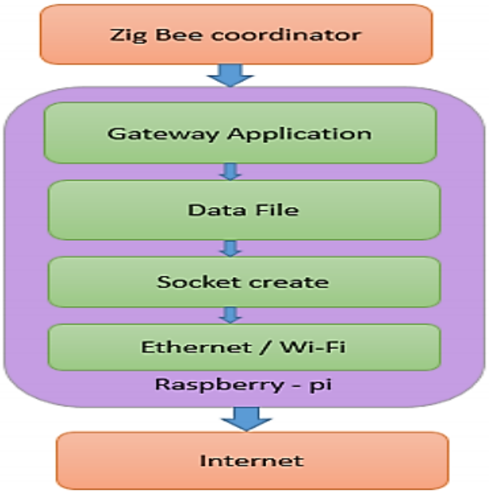


Fig 2. Functional diagram of base station.

Authors developed as system for smart agriculture. Farm area, sensor node communicates with R-Pi via Wi-Fi, Zig Bee and RF module to provide specific position sensor data on the R-pi. The data can be transmitted via internet cloud to web server. Functional diagram of Base station is shown in Fig. 2 [18].

Researchers proposed Particle Swarm Optimization energy efficient protocol that enhanced the life of the network, the Cluster head is chosen by capability functions based which consider the distance between nodes and base station [19]. A cluster aided Multipath Routing protocol is proposed by the researchers which distributed the area of interest into zones and allocate one cluster head for every cluster. And non-cluster nodes have assumed the tradeoff method for residual energy assessment between itself and nearest nodes and make decision. The authors argued that the proposed protocol reduce energy consumption because of random selection of Cluster head based on residual energies. Additionally, this protocol also corrects the tuning factors to the sink node, including remaining energy, node degree, and distance. Nonetheless, with many advantages, due to its energy measurement and random selection of CH in the network this protocol has a substantial delay [20].

Authors proposed the base technique of the Distributed Unequal Size Optimize Cluster to solve the CH load balancing problem. The BS elects the CH node based on an energy point, as well as the distance from BS, according to the protocol. The CH close the BS selects the least number of sensor nodes compared to the CH which is distant from the BS during the formation of cluster [21].

III. Motivation

Most of the research mainly focused on energy efficient routing that explores the gaps that motivate for the creation the creation of research problem. On the basis on literature study. It is found that CH is heavily responsible for directly transmitting cluster data to the BS. The CH that sends out data directly to the base station uses extra energy. Cluster head (CH) that is far away from the base station need more energy to transmit cluster data in a single hop to the base station. As a result, these problems contribute to the early depletion of cluster heads that are far-away from the bases station (BS). In majority of the protocol such LEACH in Agriculture [12]. Cluster head transfer data directly to the base station Consequently, irregular load distribution among cluster heads opted to quickly exhaust their energy which leads to disturbance of the process of data dissemination and as well shorten the network life.

IV. IoT based Agriculture

IoT nodes with limited processing power, memory, battery life and small in size are distributed in cluster farm. Due to the limited battery power IoT network needs to extend the lifetime of the system because it consumes low energy than WSN. A scheme based on clustering adapts energy use by equilibrating all nodes into a cluster head. In this research we have taken basic idea from LEACH protocol to improve the system performance, reduce energy consumption and increase the lifetime of IoT network and propose a new scheme named as IoT based Agriculture. The aims of this design are to collect and process information from different IoT nodes deployed in cluster farm and transfer data to base station (BS) for further analysis and decision making to improve crop yield. Schematic diagram is shown in Fig. 3 and assumption for mathematical model is presented in Section A.

A. *Assumption for Simulation*

Mathematical model assumptions are demonstrated below.

- In the selection farm IoT nodes are distributed randomly to ensure equal distribution.
- All the IoT nodes send hello messages with their local information to BS.
- By taking optimum values the initial number of clusters is fixed and continues to vary with node density once the node begins to die; the smaller clusters transform into larger ones.
- The BS is aggregation destination with very less power constraints and improved computation capabilities.

To obtain a satisfactory signal to noise ratio (SNR) a first order radio energy dissipation model from (LEACH) [22] is utilized to transfer a bit message over a distance d.

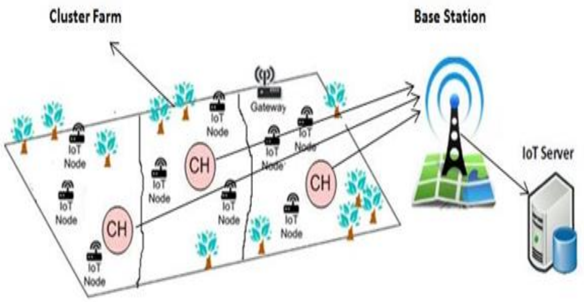


Fig 3. Schematic diagram for IoT based Agriculture.

B. Initilization Phase

For purpose of research a farm having an area of $500 \times 500\,\mathrm{m}^2$ is selected and further divided into clusters and IoT nodes are randomly deployed in different clusters using random topology. Based on their deployment and scale this research use a variety of IoT nodes that perform several jobs such as monitoring of soil moisture, fertilizers, pests / diseases and the effects of climate change. All IoT nodes deployed in the field are distributed into clusters so that all three IoT nodes are contained in each cluster. One cluster node does not interact with another cluster node but they interact with their head node only. Cluster head nodes of each cluster share data with the sink and sink with the BS.

C. First Order Radio Model

This research work takes on a first order radio model in which the radio dissipate $\mathrm{E_{elec}} = 50\,\mathrm{nl/bit}$ to run the transmitter or receiver circuit system and $\epsilon_{\mathrm{amp}} = 100$ pJ/bit/m for the transmit amplifier to achieve an satisfactory $\frac{\mathrm{E_b}}{\mathrm{N_0}}$ as shown in Fig. 4 and Table I. Another assumption was also made that an e^2 energy loss due to channel transmission. Therefore, to communicate a 3-bit message a distance d using radio models the radio expends.

$$E_{T_x}(m,d) = E_{T_{x-elec}}(m) + E_{T_{x-amp}}(m,d) \tag{1}$$

$$E_{T_x}(m,d) = E_{elec} * m + \epsilon_{amp} * m * d^2$$

And to receive this message the radio expands:

$$E_{R_x}(m) = E_{R_{x-elec}}(m)$$

$$E_{R_x}(m) = E_{elec} * m \tag{2}$$

Receiving a message is not a low-cost operation for these parameter values, so the protocols will try to minimize not only the transmission distances, but also the number of transmission and receiving operations for each message.

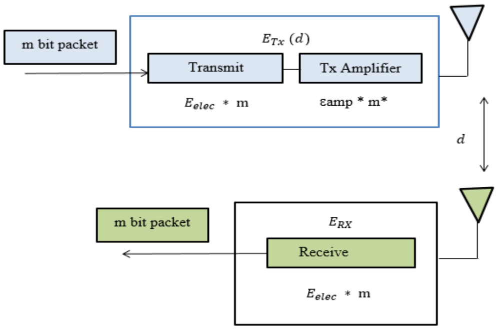


Fig 4. First Order Radio Model.

TABLE I. RADIO CHARACTRISTICS

Operations	***Energy dissipated***
Transmitter Electronics $(E_{T_{x-elec}})$	50 nJ/bit
Receiver Electronics $(E_{R_{x-elec}})$	
$(E_{T_{x-elec}} = E_{R_{x-elec}} = E_{elec})$	
Transmit Amplifier (ϵ_{amp})	100 pJ/bit/m^2

D. Clustering Mechanism

After the deployment of IoT nodes in a farm they divided into a group these groups are denoted to as clusters. A cluster can have the same types of IoT nodes or different types of nodes depending on the requirements. There is a head node in each cluster and all the IoT nodes of a specific cluster report to the similar head node and G will represent each cluster. A cluster intercluster and intracluster have two forms of interaction. The node must be told after selecting which cluster it belongs to the cluster-head node that it will be a cluster member.

- Selection of Cluster Head (CH)

CH selection considers two aspects first the ideal percentage of nodes in the network and second the history of nodes that acted as CH. Bases on the generation of random number (between 0 and 1) each n node makes the decision. In case of generated random number is less than the threshold value (Tn) the corresponding node will be CH for that round.

$$T_{(n)} = \begin{cases} \frac{\mathrm{P}}{1-\mathrm{Px}(\,\mathrm{r\,mod}\,\frac{1}{\mathrm{P}})} & \mathrm{n} \in \mathrm{G} \\ 0 & otherwise \end{cases} \tag{3}$$

As p is the suitable percentage of CH the number of rounds is r and the set of nodes is G, which in the last 1/p rounds was not CH. Unable to become cluster heads again for p rounds nodes that were cluster heads. That node then has a 1/P chance in every round to become a cluster head. In the advertising phase the Cluster heads intimate their neighbors with an advertising package that they have chosen as Cluster heads. Non cluster head nodes choose the advertising packet with the strongest signal strength received and each non cluster head node decides to join cluster upon receiving the CH broadcast. The decision can be based among other factors on the strength of signal CH broadcast message to start data transmission schedule.

- Data Transmission

Data transmission schedule start when the clusters are created and fixed. The nodes that were assumed always have data to transfer during their allocated transmission time to the cluster head. The minimum amount of energy is required for this transmission. To reduce energy dissipation in these nodes the radio member node can be switched off up to allocation of transmission time. In order to receive complete data, the cluster head node must hold its receiver on and when complete information is received then cluster head node performs function of signal processing to compress all information into a one signal. Suppose if the information is audio or seismic to produce a composite signal then CH will radiate the separate signals if it transmits composite signal to the BS then this transmission will be of high-energy because the BS is placed far away from the farm. The next round start after definite date and each node decide if it should be a CH for this round and broadcast this information.

E. Routing phase

IoT-based agricultural network consists of various types of long-range and short-range communications networks. Many IoT network technologies contribute to the design of sensors and tools for crop or field monitoring. Communication protocols are the foundation and implementations of the IoT agricultural network system and used throughout the network to exchange all agricultural data or information.

This research work proposes a new three phases IoT clustering protocol (IoT based Agriculture) for efficient usage of energy for data routing in IoT nodes which have lengthier network lifetime. In first phase IoT nodes collet data and send to their respective cluster head nodes while in second phase selected CHs receive information from their nominated nodes and send to sink while in 3rd phase received filtered information send to BS by sink as shown in Fig. 5. Routing mechanism is presented in “equation (4)”.

$$T_{(n)} = f_{(x)} = \left\{ \quad \frac{\mathrm{P}}{1-\mathrm{P}\left(\mathrm{r} \bmod \frac{\mathrm{d}}{\mathrm{P}}\right)} \quad \mathrm{x}\ \frac{E_{residual}}{E_{initial}}\ m_{opt} \text{ for all } \in\ \forall \right. \mathrm{G} \tag{4}$$

Wherever $\mathrm{E}_{\mathrm{residual}}$ is node level remaining energy and where $\mathrm{E}_{\mathrm{initial}}$ the initial is level energy assigned. Therefore, the optimal number of clusters $\mathrm{m}_{\mathrm{opt}}$ could be written as

$$m_{opt} = \sqrt{\frac{E_{fs}}{E_{amp}l^4(2m-1)E_0 - mE_{DA}} X} \tag{5}$$

In this equation the diameter of network is represented by X whereas E_0 is the initial source of energy for each node. For the current round when cluster heads are chosen, they communicate their CH selection message to other nodes in the same clusters then non cluster head nodes examine the message signal strength and decide the cluster heads to enter. After that the cluster head broadcasts timetables (Schedule) for its member nodes to transmit data in different time slots to prevent data collision. Then the process goes on for the rest of the rounds until all the nodes in the network consume all their energy.

We may adopt the following model for transmission and reception purposes as an extension of the first order radio model given in “equation (6)” that is utilized to calculate the energy consumed by each IoT node deployed in the network to transmit (ETX) and receive (ERX) packet size 1 bits over distance d.

$$E_{TX} = f(x) = \begin{cases} m * \left(E_{elec} + \epsilon_{fs} * d^2\right),\ d < d_o \\ m * \left(E_{elec+\epsilon_{mp}} * d^4\right), d \geq d_o \end{cases} \tag{6}$$

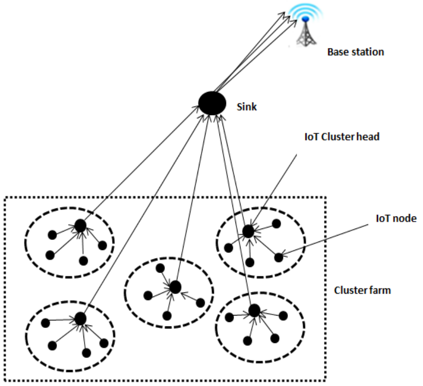


Fig 5. IoT based Agriculture Clustering Scheme.

The distance threshold d_o is the normal transmission range of IoT node. E_{elec} And ϵ_{fs} are energy dissipation to route the radio and free space model of transmitter amplifier having values 50 nJ/bit and 10 pJ/bit/m^2 accordingly, m is the data packet size and ϵ_{mp} is the multi path model of transmitter amplifier and having its value is 0.0013 pJ/bit/m4. Therefore receiving energy E_{RX} can be calculated as

$$E_{RX} = m * E_{elec} \tag{7}$$

As mentioned earlier all IoT nodes are deployed in the cluster farm randomly. The IoT nodes deployed in the cluster farm are similar having same function and same initial energy with limited power supply that cannot be re-energized or changed. Let’s assume the sink node is deployed outside the cluster farm with no energy constraints. IoT nodes which are deployed in the cluster farm are well aware about the location of the sink node and similarly about their own location and are unable to change their location. Based on the received signal strength IoT nodes make decision about the distance between one another. Depending on signal strength IoT nodes convey the information about their location and energy hop by hop to sink in the initial round. IoT nodes are attached to the grid in the same way. Those nodes which are chosen as grid head perform their allotted role while pausing sensing role. In the cluster farm other IoT nodes begin the exchange of control packets for CH or selection of first head after joining the second phase.

All IoT nodes share through the control packet their residual energy and location information. The procedure of fuzzification of shared parameters by IoT nodes starts distribution. A CH is chosen to be the node with higher residual energy and lower Euclidean distance to sink and GH. We find the Euclidean distance between any two nodes a and b from the next two dimensions of the Euclidean distance as:

$$\mathrm{d}\ (a, b) = \sqrt{(x_2 - x_1) + (y_2 - y_1})} \tag{8}$$

Euclidean distance formula where x_1 and x_2 are the width dimensions, y_1 and y_2 are length dimensions of IoT nodes a and b correspondingly. Those IoT nodes which have cluster head (CH) functions capture, compress, and transmit data from nodes to GH which is turn forward them towards sink. Whereas the other nodes in the cluster farm feel the climate, collect data and move it on to their CH. There is no need to repeat process of selecting cluster head and exchanging control packets after every round. The main aim of the method proposed is to reduce unnecessary operations in each round and to save energy to nodes. “Equation (9)” shows the Objective feature of the proposed mechanism.

$$Min \sum_{r=1}^{r=max} \mathrm{WE}_{\mathrm{consumed}}\ (\mathrm{r})\ \forall_{\mathrm{r}} \in \mathrm{R} \tag{9}$$

$\mathrm{WE}_{\mathrm{consumed}}$ is calculated using “equation (5)”

$$\mathrm{WE}_{\mathrm{consumed}} = \sum_{i=1}^{N} \mathrm{l}_0\ (\mathrm{i})\ \mathrm{x}\ (\ \mathrm{E}_{\mathrm{TX}}\text{Control packet} + \mathrm{E}_{\mathrm{RX}}\ \text{Control packet}) \tag{10}$$

The quantity of energy consumed per round within the interchange of control packets between all deployed IoT nodes n for CHs selection is considered as $\mathrm{WE}_{\mathrm{consumed}}$ and the maximal distance between two adjacent nodes is d_0. The energy consumption in transmitting and receiving control packets is assumed as E_{TX} ControlPacket and E_{RX} ControlPacket. Their values can be determined using "equations (1) and (2)" as m is equal to the size of the control packet.

In order to achieve the objective function defined in "equation (4)" we divide the total node energy into diverse equal portions and named as energy levels (EL) which can be calculated from "equation (6)".

$$EL = Eo/TL \tag{11}$$

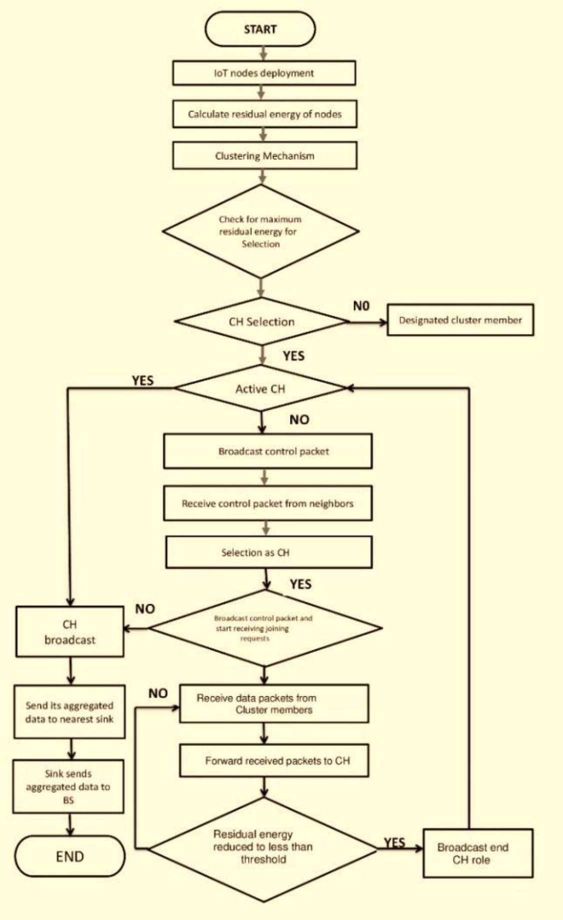

START
IoT nodes deployment
Calculate residual energy of nodes
Clustering Mechanism
Check for maximum residual energy for Selection
CH Selection
NO
Designated cluster member
YES
YES
Active CH
NO
Broadcast control packet
Receive control packet from neighbors
Selection as CH
YES
NO
CH broadcast
Broadcast control packet and start receiving joining requests
Send its aggregated data to nearest sink
NO
Receive data packets from Cluster members
Forward received packets to CH
Sink sends aggregated data to BS
END
Residual energy reduced to less than threshold
YES
Broadcast end CH role

Fig 6. Flow Chart for the Presented Scheme.

Eo is denoted as initial static node energy and TL is represented as total energy level and depends on the level of energy consumed by IoT node. The value of TL depends on the frequency of energy consumption by a node, density of the network, and data packet size and it is inversely proportional to EL. When the value of TL is low then the value of EL will be high. The value of TL cannot be selected as very low due to trade-off among all IoT nodes. IoT node chosen as CH remains as CH if the value of EL does not reduce the residual energy and there is no CH reselection phase but when the residual energy of CH reduced by EL then a control packet announcement broadcasts the end of its selection as CH. Thus, the new process of selecting CH begins a transmission of data to CHs takes place. For energy efficiency only the transmitting node remains active and all other nodes in the cluster switch off. The function of CH starts when all IoT nodes in the cluster farm have completed data transmission. The CH received data and starts aggregation to eliminate any duplication and then wrapped data for equal usage of bandwidth as much as possible.

The CH collects and then aggregates the data to eliminate any duplication and wrapping for bandwidth equal usage as much as possible. Then the CHs forward the data in single-hop to the sink and then to the BS. The whole cycle is presented in steady state phase in the form of flow chart.

- Steady State phase

Flow Chart for the Presented Scheme is shown in Fig. 6.

V. Simulation Results and Discussion

IoT is bigger than that the Wireless Sensor Network. WSN is an innovation that is frequently used within an IoT system. A costly collection of sensors as in a mesh network can be used to collect information individually and send information in an IoT system via switch to the web. In an IoT network most of the nodes directly send data to the web. Such as a node might be used for the measurement of the soil temperature. In this scenario the data will be frequently or infrequently send to the web directly. Where the information can be handled by a server and needs to be interpreted on a front-end interface.

Table II defines the model parameters considered for the network model for MATLAB simulation. The size of the packet is 200 bits. 100 nodes are deployed randomly with sink and BS is positioned outside the field.

- Network stability period

From network start up to the death of the first node is called network stability period. It is shown in Fig. 7 and Table III that stability period for IoT based Agriculture is much greater than that of LEACH in Agriculture. The reason of improvement is transmission of non-continuous data. In IoT-based agriculture data will only be transmitted if there is a difference between the current sensed value and the previously sensed value. The first node of LEACH in Agriculture dies at 168 arounds whereas IoT based Agriculture first node dies after 463 rounds hence IoT based Agriculture shows 23% improvement in network stability.

TABLE II. Simulation Parameters

Parameters	***Values***
System diameter	$500 \times 500m^2$
Total number of IoT nodes (n)	100
Packet Size (m)	200 bits
Initial energy of node	0.9 J
Required percentage of CH selection (p)	0.1 J
Transmitter energy	50×10^{-8}nJ/bit/m^2
Receiver energy	50×10^{-8}nJ/bit/m^2
Trasnmit amplifier d < do	10×10^{-11}nJ/bit/m^2
Transmit amplifier d > d_0	0.0013 x 10^{-8}nJ/bit/m^2
Data aggregation energy cost	5×10^{-11}pJ/bit

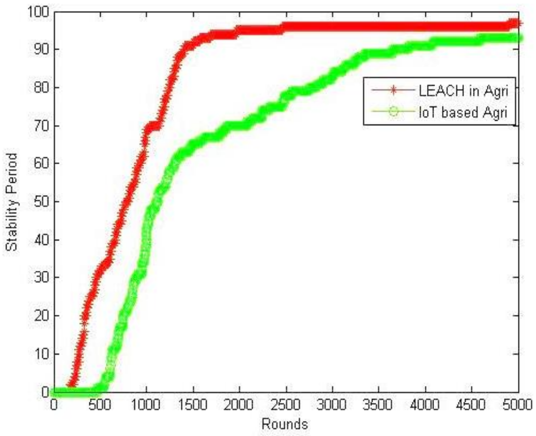


Fig 7. Stability Period of LEACH in Agriculture and IoT based Agriculture.

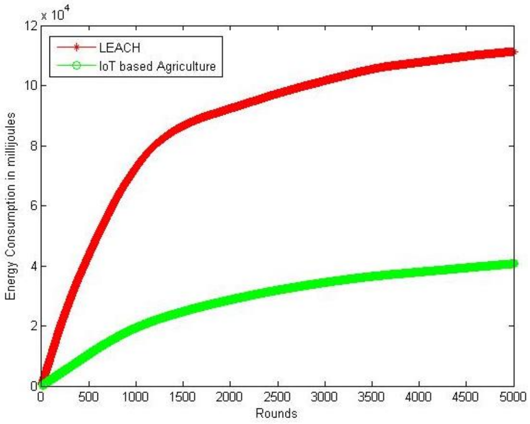


Fig 8. Energy Consumption of LEACH in Agriculture and IoT based Agriculture.

- Energy consumption

Energy consumption is a heart of any network. It is the net energy consumed by nodes during network operation. If network uses low energy, it will have longer life and if it uses high energy, it will have limited life time so that energy consumption has an effect on the life of the network. WSN is an ad hoc network and cannot send data directly to the internet due to which it consumed high energy whereas IoT is a one hop transmission and send data directly to the internet therefore IoT network consumed less energy than that of WSN network. In IoT based Agriculture initial energy is equivalent to 0.9 J and simulation results are shown in Fig. 8 and Table IV the energy consumption of IoT based Agriculture is 68% less than that of LEACH in Agriculture which may prolong the life of network.

- Network life

It is the time till the first node energy runs out it is an important performance metric. In IoT network all of the nodes without delay send their records to the internet such as a node might also be used to monitor the temperature of the soil. In this setup the information will be straight away or periodically dispatched at once to the web. Whereas in Wireless sensor network setup there is no direct communication to the internet. In its place the several sensors linked to router which further transfers data to the internet. Therefore, the IoT network has longer life than that of WSN network. For each round, LEACH in Agriculture assumes that CHs dissipates the same energy that results in inefficient CH selection and affects the lifespan of the network. IoT based agriculture selects CHs considering the residual energy of nodes and the optimal number of clusters together, thereby increasing the lifetime of the network to more rounds and showing an improvement of 112 percent as shown Fig. 9 and Table V.

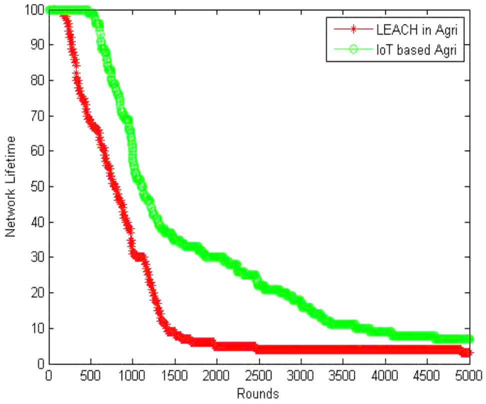


Fig 9. Network Life of LEACH in Agriculture and IoT based Agriculture.

TABLE III. NETOWRK STABILITY OF LEACH IN AGRICULURE VS IOT BASED AGRICULTURE

Scheme name	**Rounds**											
	500	**1000**	**1500**	**2000**	**2500**	**3000**	**3500**	**4000**	**4500**	**5000**	**Average**	**Improvement %age**
LEACH in Agriculture	32	69	91	95	96	96	96	96	96	97	86.4	100
IoT based Agriculture	1	42	65	70	78	83	89	91	91	93	70.3	123

TABLE IV. ENERGY CONSUMPTION OF LEACH IN AGRICULURE VS IOT BASED AGRICULTURE

Scheme name		**Rounds**										
	500	**1000**	**1500**	**2000**	**2500**	**3000**	**3500**	**4000**	**4500**	**5000**	**Average**	**Improvement %age**
LEACH in Agriculture	4.35	7.27	8.65	9.23	9.73	10.11	10.5	10.71	10.91	11.12	9.25	100
IoT based Agriculture	1.05	1.09	2.47	2.86	3.19	3.44	3.64	3.79	3.93	4.07	2,95	31.89

TABLE V. NETWORK LIFE TIME OF LEACH IN AGRICULURE VS IOT BASED AGRICULTURE

Scheme name		**Rounds**										
	500	**1000**	**1500**	**2000**	**2500**	**3000**	**3500**	**4000**	**4500**	**5000**	**Average**	**Improvement %age**
LEACH in Agriculture	68	31	8	5	4	8	4	4	4	3	13.9	100
IoT based Agriculture	99	58	35	30	22	17	11	9	8	7	29.6	212

VI. CONCLUSION

In the agricultural sector, the Internet of Things IoT is one of the emerging technologies. IoT nodes senses soil and crop physical and environmental conditions and sends the data to towards the sink BS which reduces load between the CH. We performed the simulation to evaluate the efficiency of the proposed IoT-based Agriculture scheme for various criteria including network stability, energy consumption and network life. The results showed that our scheme is better network reliability, energy consumption and network life efficiency than LEACH protocol in agriculture. For a feasible routing route, most routing protocols don't allow load balancing. This work particularly on the cluster head (CH) improves the load between the IoT nodes. The scheme proposed selects and rotates the CH close to the cluster's energy centroid position. Therefore, each CH selects the gateway node for multihopping itself and other CH data towards the sink and BS which reduces load between the CH. In our future research, we will integrate blockchain with IoT and will compare the performance results with IoT based Agriculture scheme and will develop a smart model for agricultural environment monitoring.

REFERENCES

[1] Fathy, Yasmin, Payam Barnaghi, and Rahim Tafazolli. "Large-scale indexing, discovery, and ranking for the Internet of Things (IoT)." *ACM Computing Surveys (CSUR)* 51, no. 2 (2018): 1-53.

[2] Eastwood, Callum, Laurens Klerkx, Margaret Ayre, and B. Dela Rue. "Managing socio-ethical challenges in the development of smart farming: from a fragmented to a comprehensive approach for responsible research and innovation." Journal of Agricultural and Environmental Ethics32, no. 5-6 (2019): 741-768.

[3] Ray, Partha Pratim. "A survey on Internet of Things architectures." Journal of King Saud University-Computer and Information Sciences 30, no. 3 (2018): 291-319.

[4] Khan, Fazeel Ahmed, Adamu Abubakar, Marwan Mahmoud, Mahmoud Ahmad Al-Khasawneh, and Ala Abdulsalam Alarood. "Cotton Crop Cultivation Oriented Semantic Framework Based on IoT Smart Farming Application." International Journal of Engineering and Advanced Technology 8, no. 3 (2019): 480-484.

[5] Reynolds, Daniel, Joshua Ball, Alan Bauer, Robert Davey, Simon Griffiths, and Ji Zhou. "CropSight: a scalable and open-source information management system for distributed plant phenotyping and IoT-based crop management." Gigascience8, no. 3 (2019): giz009.

[6] Srivastava, Vyoma, K. K. Aggarwal, and Abhay Kumar Srivastava. A Revisit to Clustering Techniques with its Application in Agriculture Sector. No. 793. EasyChair, 2019.

[7] Prathibha, S. R., Anupama Hongal, and M. P. Jyothi. "IoT based monitoring system in smart agriculture." In 2017 International Conference on Recent Advances in Electronics and Communication Technology (ICRAECT), pp. 81-84. IEEE, 2017.

[8] Sholla, Sahil, Sukhkirandeep Kaur, Gh Rasool Begh, Roohie Naaz Mir, and M. Ahsan Chishti. "Clustering Internet of Things: A Review." Journal of Science and Technology: Issue on Information and Communications Technology 3, no. 2 (2017): 21-27.

[9] Awuor, Fredrick, George Raburu, ArvinLucy A. Onditi, and Dorothy Rambim. "Building e-agriculture framework in Kenya." (2016).

[10] Adriano, José D., Yara CT Mendes, Guilherme AB Marcondes, Vasco Furtado, and Joel JPC Rodrigues. "An IoT Sensor Mote for Precision Agriculture with Several MAC Layer Protocols Support." In 2018 International Conference on Information and Communication Technology Convergence (ICTC), pp. 684-688. IEEE, 2018.

[11] Sahitya, Gadikota, Narayanam Balaji, Challa Dhanuanjaya Naidu, and S. Abinaya. "Designing a wireless sensor network for precision agriculture using ZigBee." In 2017 IEEE 7th International Advance Computing Conference (IACC), IEEE, 2017.

[12] Aung, Than Htike, Su Su Yi Mon, Chaw Myat Nwe, Zaw Min Naing, and HLa Myo Tun "Implementation Of The Precision Agriculture Using LEACH Protocol Of Wireless Sensor Network."

[13] Gupta, Suneet Kumar, Pratyay Kuila, and Prasanta K. Jana. "Energy efficient multipath routing for wireless sensor networks: A genetic algorithm approach." In 2016 international conference on advances in computing, communications and informatics (ICACCI), pp. 1735-1740. IEEE, 2016.

[14] Khomami, Mohammad Mehdi Daliri, Alireza Rezvanian, and Mohammad Reza Meybodi. "Distributed learning automata-based algorithm for community detection in complex networks." International Journal of Modern Physics B 30, no. 8 (2016): 1650042.

[15] ZHAO, Aqun, and Qi ZHAO. "A New Routing Algorithm for Multi-path Transmission." Przegląd Elektrotechniczny 89, no. 1b (2013): 211-213.

[16] Wang, Tian, Lei Qiu, Guangquan Xu, Arun Kumar Sangaiah, and Anfeng Liu. "Energy-efficient and trustworthy data collection protocol based on mobile fog computing in Internet of Things." IEEE Transactions on Industrial Informatics (2019).

[17] Umer, Tariq, Mubashir Husain Rehmani, Ahmed E. Kamal, and Lyudmila Mihaylova. "Information and resource management systems for Internet of Things: Energy management, communication protocols and future applications." (2019): 1021-1027.

[18] Vyas, Dharti, Amol Borole, and Shikha Singh. "Smart agriculture monitoring and data acqusition system." International Research Journal of Engineering and Technology 3 (2016): 1823-1826.

[19] Qureshi, Kashif Naseer, Muhammad Umair Bashir, Jaime Lloret, and Antonio Leon. "Optimized Cluster-Based Dynamic Energy-Aware Routing Protocol for Wireless Sensor Networks in Agriculture Precision." Journal of Sensors 2020 (2020).

[20] Sajwan, Mohit, Devashish Gosain, and Ajay K. Sharma. "CAMP: cluster aided multi-path routing protocol for wireless sensor networks." Wireless Networks 25, no. 5 (2019): 2603-2620.

[21] Mishra, Kaushik, and Santosh Majhi. "A State-of-Art on Cloud Load Balancing Algorithms." International Journal of Computing and Digital Systems 9, no. 2 (2020): 201-220.

[22] Heinzelman, Wendi Rabiner, Anantha Chandrakasan, and Hari Balakrishnan. "Energy-efficient communication protocol for wireless microsensor networks." In Proceedings of the 33rd annual Hawaii international conference on system sciences, pp. 10-pp. IEEE, 2000.

Video Genre Classification using Convolutional Recurrent Neural Networks

Dr K Prasanna Lakshmi[1]
Professor and Head, Information Technology Department, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India

Mihir Solanki[2], Jyothi Swaroop Dara[3]
Information Technology Department
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India

Avinash Bhargav Kompalli[4]
Department of CSE
SRM University,
Chennai, India

***Abstract*—A wide amount of media in the internet is in the form of video files which have different formats and encodings. Easy identification and sorting of videos becomes a mammoth task if done manually. With an ever-increasing demand for video streaming and download, the Video Classification problem is brought into foresight for managing such large and unstructured data over the internet and locally. We present a solution for classifying videos into genres and locality by training a Convolutional Recurrent Neural Network. It involves feature extraction from video files in the form of frames and audio. The Neural Networks makes a suitable prediction. The final output layer will place the video in a certain genre. This problem could be applied to a vast number of applications including but not limited to search optimization, grouping, critic reviews, piracy detection, targeted advertisements, etc. We expect our fully trained model to identify, with acceptable accuracy, any video or video clip over the internet and thus eliminate the cumbersome problem of manual video classification.**

Keywords*—Convolutional recurrent neural networks; video classification; temporal and spatial aspects; machine learning; computer vision; images classification; audio classification

I. Introduction

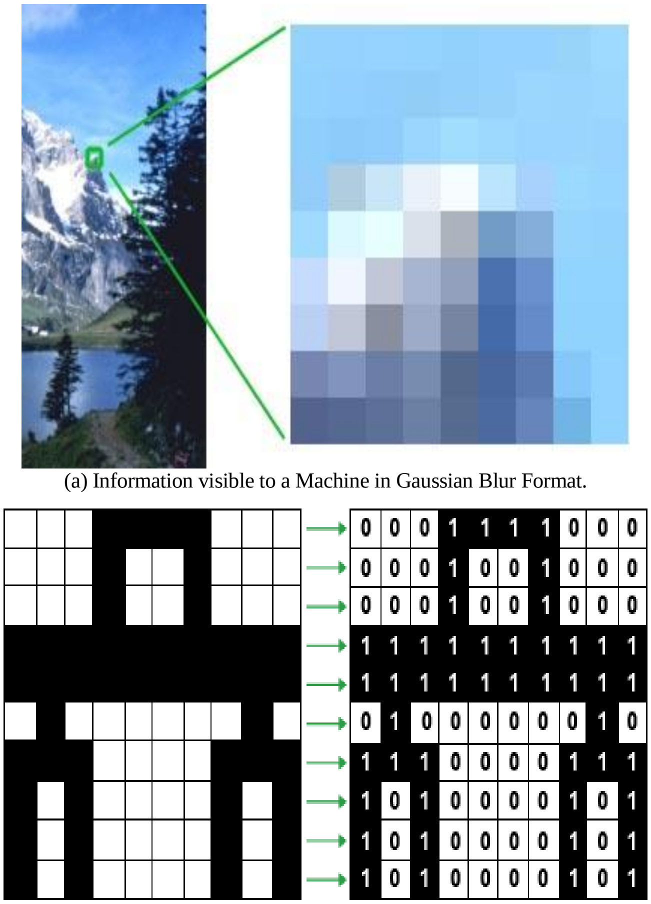
By and large all techniques used in video classification have been image based, with little consideration going into the background audio and annotations. CNN-LSTMs [1] have shown great strides in recognizing image-based video inputs and classifying them into output categories. As humans though, we not only recognize a video by its visual features, but also by the perceived audio it generates. To teach a machine to take similar features into consideration would make a lot of sense because audio plays a large role in classifying videos too. For example, an action scene in a movie will have a fast-paced audio accompanying it, a serious dialogue session will have a lot of voices and weak music notes. Also, a lot of video shot in the internet could be amateurish, with blurry images and weird camera angles.

Giving the context of audio will help the Neural Network more features to rely on while making a classification.

To a human a video is a ray of different colors striking the eyes, but computers perceive video in a completely different way from us. At the lowest level, it is a series of 1's and 0's which makes no sense to the processor except to light up a certain pixel in certain color. When we teach a Neural Network to identify videos, we are asking it to identify certain patterns in those numbers based on mathematical calculations. An image, therefore may be viewed by a machine as in Fig. 1(a) and 1(b).

The problem inherent in computer vision, in fact, the very purpose of the field, is to recover information about the world from sensory input. This can be thought about as a formula:

$$S = f(W) \tag{1}$$



(a) Information visible to a Machine in Gaussian Blur Format.

0	0	0	1	1	1	1	0	0	0
0	0	0	1	0	0	1	0	0	0
0	0	0	1	0	0	1	0	0	0
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
0	1	0	0	0	0	0	0	1	0
1	1	1	0	0	0	0	1	1	1
1	0	1	0	0	0	0	1	0	1
1	0	1	0	0	0	0	1	0	1
1	0	1	0	0	0	0	1	0	1

Fig 1. (b): Information visible to a Machine in Bitmap Format.

Our sensory information(S) is a function of the world (W) around us (1). What humans take for granted, and what the field of Computer Vision struggles to make machines do, is the reverse:

$$W = f^{-1}(S) \tag{2}$$

That is, to understand the world from sensory information (2).

Audio is different scene altogether. A common way to input audio to Machine Learning algorithms is by using a Mel spectrogram. A mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

MFC coefficients are commonly derived as follows:

1) Take the Fourier transform of (a windowed excerpt of) a signal.

2) Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.

3) Take the logs of the powers at each of the mel frequencies.

4) Take the discrete cosine transform of the list of mel log powers, as if it were a signal.

5) The MFCCs are the amplitudes of the resulting spectrum.

A popular formula to convert *f* hertz into *m* mels is in Fig. 2. Fig. 3 shows the generated Mel spectrogram of an audio file using Audacity.

Most state-of-the-art algorithms use this technique as a baseline for their inputs. Hence, we've chosen the same techniques for the inputs to our model. Combining the best of both audio and video classification techniques, we present a unique solution for the video genre classification problem using a Convolutional Recurrent Neural Network or a Convolutional Long Short-Term Memory Network.

Convolutional Neural Networks have been the best at spatial feature extraction and classification problems for images. Some popular examples are ImageNet [2], MobileNet [3], Inception [3], and Google's WaveNet [4]. Feature extraction from a single frame may be straightforward, however a video is a sequence of frames, and every frame is important. For example, we cannot recognize an action of say eating a bowl of cereal, until we have seen a person putting a spoon into the bowl and then into his mouth. Similarly, we must teach a machine to not only look at one frame, but a sequence of frames, to grasp the context of the video. The same analogy can be applied to audio. Hearing a single beat will not help us identify the genre of a song. Only when we hear it for a few seconds are we be able to identify its tempo, the instruments used and its theme. This is where Recurrent Neural Networks come into play. A recurrent neural network (RNN) is a class of artificial neural network where connections between nodes form a directed graph along a temporal sequence. So RNNs can not only understand features at a single timestep, but also remember features from previous timesteps, making them best suited for solving temporal region problems. Long short-term memory (LSTM) [20] follow the RNN architecture and have shown great promise in the video classification problem.

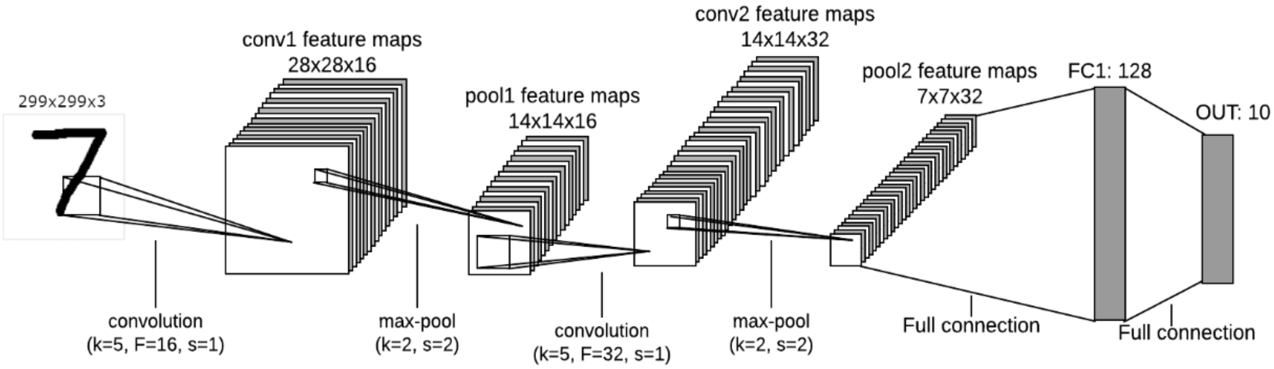


Fig 2. The Convolutional Neural Network Architecture.

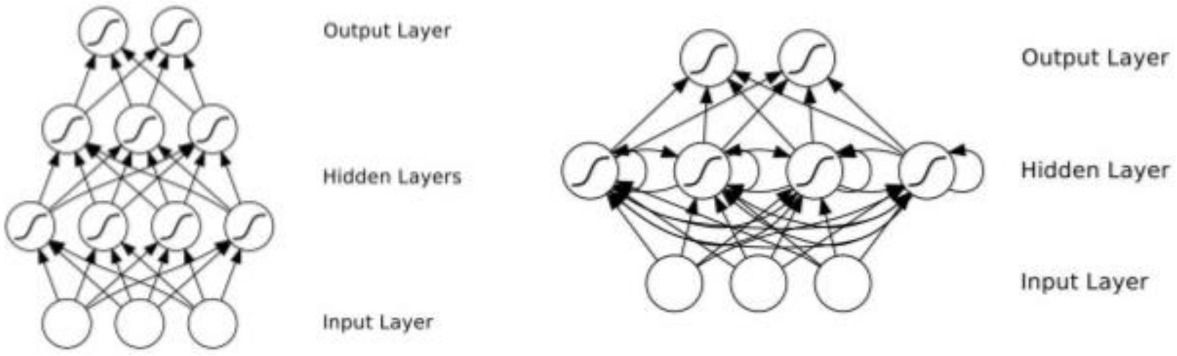


Fig 3. Simplified Representation of a Convolutional Neural Network and a Recurrent Neural Network.

The CNN output can be taken in two methods. One is we take the output from the SoftMax layer, which includes the predictions the CNN has made. The other method is to use the output from the pool layer, which leaves the output prediction to the RNN. We have tried both methods for this paper and they are explained in detail later.

II. Genre in Videos

A genre for a video specifies a certain expectation about the video. Genres in real world videos are neither specific nor implicit but tend to be overlapping. Also, they vary from person to person as perspective matters. In such a case, defining specific boundaries for genres tends to become difficult. For example, there is a very thin line between the genres Drama and Thriller, and many film critics argue for the same. To define audio into genres has a different shortcoming. Audio Classification is usually multi-label, because they tend to be a mixture of multiple tastes and themes. Background music in modern movies tend to be a mixture of both classical and contemporary, two very different genres if seen separately. Period movies and biographies today are examples for the above.

Therefore, to define the genres for a classifier, we must ensure that we remove the maximum conflicts that occur in genre identification. Hence, we have chosen 6 genres which we can safely say are non-overlapping and mutually exclusive, even if based on different perspectives. The genres we chose are: Action, Animation, Horror, Romance, Sports and Science Fiction. This ensures that our model does not form any tight assumptions about one genre and is also flexible and open for new genres in the future.

III. Related Work in Video Classification

A. *Truly Multi-modal YouTube-8M Video Classification with Video, Audio, and Text [5]*

Zhe Wang, Kingsley Kuan, Mathieu Ravaut and others[5] present a novel way in Video classification by using multi-modal features from audio, video and text.Their algorithm classifies the YouTube 8M dataset, which is a collection of over 0.7 million YouTube videos , each labelled automatically, without human curation. The challenge involves classifying an imbalanced dataset based on user generated video content on YouTube. They used TextCNN for titles and Random Forest and max pooling for frame classification. Their research showed that the inclusion of text yielded state-of-the-art results, e.g. 86.7% GAP on the YouTube-8M-Text validation dataset.

B. *Large Scale Video Classification using both visual and audio Features on YouTube-8M Dataset [6]*

Emma An, Anqi Ji and Edward Ng. [6] presented a solution for the YouTube-8M challenge by considering both audio and video features. They used a Convolutional Neural Network to classify videos into their 4716 classes. Their model used a Mixture of experts (MoE) to receive 3 inputs, video level features only, audio level features only and a concatenation of both audio and video features. They applied a dense layer, followed by a rule activation layer in their model. Taking the softmax function output, they achieve an AvgHit of 0.84, Avg PERR (average precision at equal recall rate) of 0.709, and mAP (mean average precision) of 0.415 compared to the best performing baseline.

C. *Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification [7]*

Ali Diba, Mohsen Fayyaz, Vivek Sharma and others [7] introduced new 3D convolutional neural network architectures for video classification named DenseNet3D and T3D.They introduced a new temporal layer that models variable temporal convolution kernel depths, embedding this new temporal layer in their proposed 3D CNN, thus extend the DenseNet architecture - which normally is 2D - with 3D filters and pooling kernels. Their research mainly dealt with action recognition in videos, using the Sports-1M, HMDB and UCF101 datasets. They beat algorithms trained in multi-GPU setup for days by removing bottlenecks in the knowledge gained by 2D ConvNets.

D. *Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks [8]*

Pouya Bashivan, Irina Rish, M. Yeasin, and Noel Codella [8], applied Deep Recurrent-Convolutional Neural Networks in classifying electroencephalogram data. Electroencephalography (EEG) is an electrophysiological monitoring method to record electrical activity of the brain. It is typically non-invasive, with the electrodes placed along the scalp, although invasive electrodes are sometimes used, as in electrocorticography. EEG measures voltage fluctuations resulting from ionic current within the neurons of the brain. By training their model, they were successful in demonstrating significant improvements in classification accuracy over current state-of-the-art approaches in this field. A similar CNN-LSTM [21] is used in this paper.

Although a hot topic in Computer Vision, surprisingly less research has been done in the category of genre identification in videos. Most of the state-of-the-art research has been done on image recognition and on solely visual features. The challenges posed for such a classification are noisy data, huge computational costs, large size of datasets and locality/copyright of videos. Some of the limitations of the above papers are:

- They classify videos into categories of fixed actions, which are very specific.
- Most researchers use the YouTube-8M [9] dataset, which is a highly imbalanced dataset and contains very generic categories.
- They rely solely on visual features, ignoring a large amount of audio data.
- They rarely consider temporal space, relying on only spatial features, which bottlenecks most classification attempts.

Through this paper, we attempt to outline and demonstrate methods to improve video classification by fixing most of the limitation mentioned above. Our research is solely academic and is meant to spark interest into the genre-classification problem and its current limitations.

IV. Data Gathering

There are certain limitations while using existing popular video datasets like YouTube-8M, HMDB [10], UCF101 [11] for genre identification problems, mainly because their labels are not categorized into movie genres. For examples categories like playing sports are placed into human actions, which should instead be classified into a sports genre by our model. Hence, we had to do a lot of manual data cleaning to get our training set.

This paper presents the work which are used in parts or in their entirety as follows:

- The UCF101 dataset
- The Hollywood2 [12] dataset
- The HMDB dataset
- The YouTube-8M dataset

We choose video from these pre-labelled datasets and classified them into folders representing our six genres. For example, videos of punching, fighting and explosions went into the action folder, cases of hauntings and paranormal scenes went into the horror folder and so on. For animation however, we had to take an entirely different approach since there is a dearth of freely available animation videos for research purposes on the internet. We resorted to manually downloading clips from public domain websites [13] [14].

Most of the animated videos found online were old hand-drawn ones, but we were able to secure some modern 3D animation from the blender.org foundation and other open sources.

The compiled dataset now consisted of 39GB of videos, each separated into folders whose names displayed their labels.

V. Data Cleaning and Preprocessing

A movie video file is usually run at a constant 24 frames per second. If we convert an entire video file into frames, we would get 24 images for a second, which when scaled for a 2-minute video amounts to 2,880 frames. This data is a lot for a model to process and therefore we had to cut down on frame count by taking only 4 frames for each second. This limit was decided after a simple test conducted on human subjects. We split different videos into 2, 4, 6 and 8 frames per second and asked the subjects to cycle through the images and guess the action to be performed. We found that the human mind could perceive any action taking place optimally in 4 frames every second, where 2 would be difficult for slow actions and 6 and 8 would be too easy to guess. Hence, we concluded that an average of 4 frames per second is enough information for a model to recognize what is going on in a frame of time as depicted in Fig. 4. The conversion of video to frames was done with the well-known library OpenCV2 [13] written in Python3.The frames were arranged in similar folders as the videos, with folder names specifying the label.

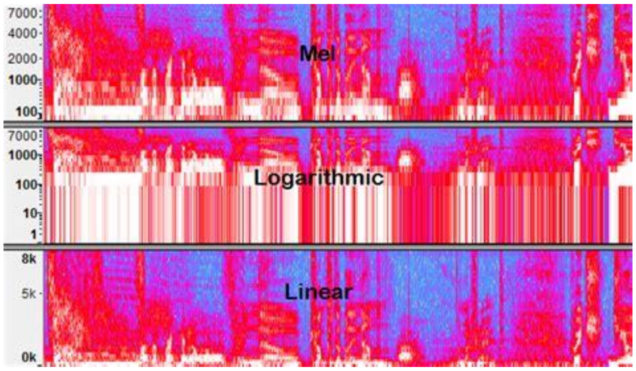


Fig 4. A visualization of different Audio (WAV File) to Frequency Graph Conversion Techniques.

For audio, we used the well-known codec FFMPEG [14]. It provides fast, efficient and lossless conversion of video files into wav files. Then each WAV file was converted into a mel Frequency Spectrogram using the Python 3 library matplotlib [15] and stored into a similar folder structure as above.

Overfitting happens when the model fits too well to the training set. It then becomes difficult for the model to generalize to new examples that were not in the training set. For example, the model recognizes specific images in the training set instead of general patterns. The training accuracy will be higher than the accuracy on the validation/test set. To prevent overfitting, we needed regular validation checks as most of our dataset consisted of specific videos. Hence, we split the set into 80/10/10 for the training, testing and validation sets respectively.

VI. The Convolutional Recurrent Neural Network Approach

A high-level architecture view of the model is shown in Fig. 3. Both audio and visual features were essentially treated as images, so they could be easily vectorized. This ensured us to categorize inputs easily as a TensorFlow/NumPy array to be given to the model.

A. *The Convolutional Neural Network*

Pouring research into the availability of state-of-the-art open-source CNNs like ImageNet, VGGNet [16], InceptionV3 [19] and others, we were able to reduce the resource intensive and repetitive task of preparing a CNN model without a baseline. It would prove more time-consuming since we needed a network that was aware of what an image was first, before it could start finding patterns. Hence, we decided on training our dataset on the pre-created models as a baseline. Inception was selected as our base CNN due to its ability of transfer learning for new classes of data as well as better accuracy for home and amateur clips. The InceptionV3 is a neural network architecture for image classification, originally published by Christian Szegedy [19], Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna[17]. This model has already been trained on a similar task for thousands of images and thus comes with an intuition for feature extraction from images.

The input to this layer comes in the form of images of size 225 x 225 x 3(width x height x channels) which are scaled accordingly using NumPy [18].

We train the session for a total of 4000 steps with the default hyper-parameters. Training checkpoints are created every 400 steps. The output will give us a retrained graph in pb format and a text file containing labels. However, we are more interested in the output of the pool and softmax layers. The layer output was taken accordingly in code and then passed on to the RNN. The reason why softmax is useful is because it converts the output of the last layer in the neural network into what is essentially a probability distribution. This gives the RNN more data to work with rather than a single 2048 vector and a class label. The advantage here is that instead of just getting a predefined label as output, we are giving our next iteration the entire data that led to its prediction of a particular label. At the end of this process we have both the vector arrays containing the features as well as the prediction probability of each class label for that vector.

B. The Recurrent Neural Network

We could choose to build our RNN either as a deeper network or as a wider network. Testing with both options, we discovered better training results while using a wider network. Another way to think about RNNs is that they have a "memory" which captures information about what has been calculated so far. In theory RNNs can make use of information in arbitrarily long sequences, but in practice they are limited to looking back only a few steps. The RNN-LSTM [21] has 2 main layers, viz. LSTM layer and the regression layer. The LSTM layer provides the temporal feature extraction that we need for the video. Denoting $*$ as elementwise multiplication and ignore bias term, LSTM calculates a hidden state ht as:

$$i_t=\sigma(x_tU_i+h_{t-1}W_i)$$

$$f_t=\sigma(x_tU_f+h_{t-1}W_f)$$

$$o_t=\sigma(x_tU_o+h_{t-1}W_o)$$

$$\sim C_t=\tanh(x_tU_g+h_{t-1}W_g)$$

$$C_t=\sigma(f_t*C_{t-1}+i_t*\sim C_t)$$

$$h_t=\tanh(C_t)*o_t \tag{3}$$

Here, *i*, *f*, *o* are called the input, forget and output gates, respectively. These gates have the exact same equations, just with different parameter matrices (*W* is the recurrent connection at the previous hidden layer and current hidden layer, *U* is the weight matrix connecting the inputs to the current hidden layer). They are called gates because the sigmoid function squashes the values of these vectors between 0 and 1, and by multiplying them element wise with another vector it defines the part of the other vector that is allowed to the next layer. The input gate defines how much of the newly computed state for the current input you want to allow to the next layer. The forget gate defines how much of the previous state you want to allow to the next layer. Finally, the output gate defines how much of the internal state you want to expose to the external network (higher layers and the next time step). All the gates have the same dimensions d_h, the size of your hidden state. ~*C* is a candidate hidden state that is computed based on the current input and the previous hidden state. *C* is the internal memory of the unit. It is a combination of the previous memory, multiplied by the forget gate, and the newly computed hidden state, multiplied by the input gate. Thus, intuitively it is a combination of how we want to combine previous memory and the new input. We could choose to ignore the old memory completely (forget gate all 0's) or ignore the newly computed state completely (input gate all 0's), but most likely we want something in between these two extremes. h_t is output hidden state, computed by multiplying the memory with the output gate. Not all of the internal memory may be relevant to the hidden state used by other units in the network.

That sequential information is preserved in the recurrent network's hidden state, which manages to span many time steps as it cascades forward to affect the processing of each new example. It is finding correlations between events separated by many moments, and these correlations are called "long-term dependencies", because an event downstream in time depends upon, and is a function of, one or more events that came before. Mathematically, the carrying forward of memory is represented as:

$$h_t=\varphi(Wx_t + Uh_{t-1}) \tag{4}$$

The hidden state at time step t is `h_t`. It is a function of the input at the same time step `x_t`, modified by a weight matrix `W` (like the one we used for feedforward nets) added to the hidden state of the previous time step `h_t-1` multiplied by its own hidden-state-to-hidden-state matrix `U`, otherwise known as a transition matrix and similar to a Markov chain. The weight matrices are filters that determine how much importance to accord to both the present input and the past hidden state. The error they generate will return via backpropagation and be used to adjust their weights until error can't go any lower.

From the output of the CNN, we group the vector sequences into 40 frames, giving us 10 seconds of information to process. The RNN has 2056 nodes and gives the output as the six classes with their probabilities. The label with the most probability assumed as the predicted class for the current frame sequence. Fig. 5 shows the architecture of our RNN.

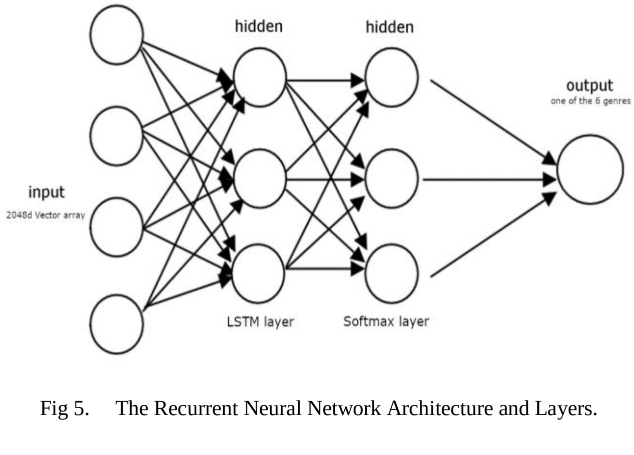


Fig 5. The Recurrent Neural Network Architecture and Layers.

VII. Training Specifications

Training video classifiers require tremendous hardware capabilities due to the size and structure of data. We decided to use the Google Cloud Platform for training our model. A Deep Learning AMI by Google was deployed on the platform and its' specifications were:

- 4x Intel XEON vCPUs
- 1x NVIDIA Tesla K80 with 12GB VRAM
- 10GB of RAM
- 100GB of fast SSD
- Debian OS

This provided us a fast, reliable, on-the-go and cost-effective solution for cloud training.

VIII. Experimental Results

With the dataset and model ready, the training took us 4 hours for the CNN part and 4 hours for the RNN part, running on the machine specified above.

The accuracy mark when we used the output from the softmax layer method, that is taking the output from the second layer, yielded 85.4%. This method gave raw data from CNN to the RNN, hence the RNN had an upper hand in making a decision. The TensorFlow log is attached in Fig. 6.

To further improve this, we used the pool layer method which took output from the third layer. This gave more computational power to the CNN and the predictions were narrowed down. This brought the accuracy mark up to 90.3%.

We then tested the model on completely unknown videos from the internet. They consisted of movies, science fiction documentaries, live sport matches and TV series. Our algorithm was able to safely classify videos by observing the temporal space in most of the cases. The shortcomings are discussed later.

To set a benchmark for our method, we trained the naïve model on the UCF101 dataset. The model was able to beat the average accuracy benchmark set on the dataset after just 3 hours of training. Table I lists the comparison of accuracies for different Video Classification methods applied on the dataset.

TABLE I. A Comparison of Various Video Classification Techniques used in the Paper

Sno	***Name***	***Accuracy***
1	ConvNet[22]	65%
2	Time distributed CNN [23]	41%
3	3D convolutional Network[24]	52.8%
4	CNN-RNN	74%
5	**CNN-LSTM – soft-max**	**85.4%**
6	**CNN-LSTM – pool**	**90.3%**

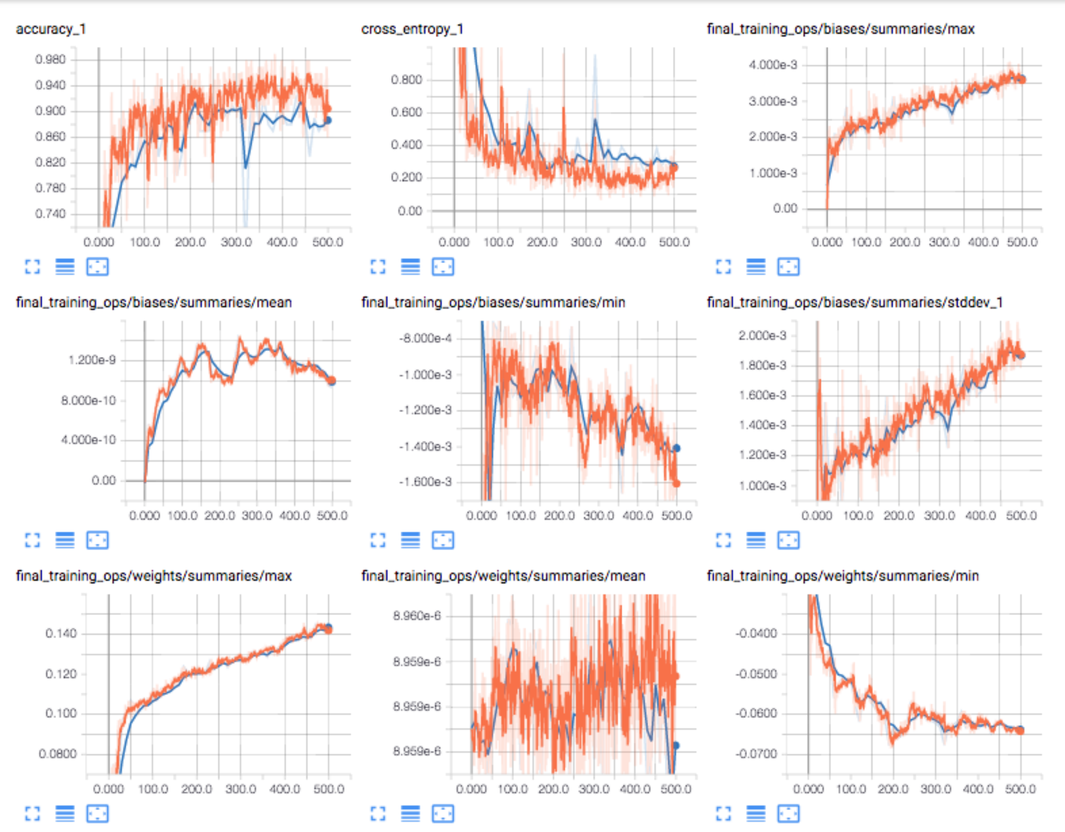


Fig 6. Tensor board Training Graphs for the CNN-LSTM Network.

IX. Conclusion and Future Enhancements

Video classification is a long open problem with tremendous possibilities for applications in the fields of medicine, entertainment, surveillance, search optimization and many others. Using only visual features has inputs leave a lot of gap for the classification methods to fill. By using audio features, we aim to fill this gap and also make a machine more intelligent while dealing with data. Video files take up a huge chunk of data stored on the internet and easy classification will always be a prime problem to be solved. The lack of proper datasets, copyright issues, video quality, etc. will always continue to be bottlenecks in the way of this problem. However, as more open-source research is made into this field, we can expect to see more efficient methods emerge which are not so computationally expensive. Our paper highlights the main shortcomings many video classifiers are plagued with, namely in using audio features and in the temporal space. Computer vision is and will be a booming field in the years to come as we move to autonomous machines and robots. Video feeds are the best input we can give to these intelligent machines.

However, there is still a long way to go before we can completely trust machines to make prediction on genres. In our testing we found two interesting cases where the model classified a certain genre wrong. In the first case; the input video we gave was from a horror movie, where a ghost is walking vertically on a tree trunk. The model continued to classify the video as action despite there being clear elements of horror present in the scene. Another case is highlighted in the genre of Romance, where due to the lack of lighting and the expressions of the actress, the model thinks the genre is horror. Such false classifications will always arrive as long as machines are ignorant about a lot of other features like human emotions and the technicalities involved in the direction of a movie. To bridge this gap would be a major step in building an AI critic, who could not only classify movies, but also judge their effectiveness and themes.

Every project is at any stage a work in progress, since we cannot achieve a perfect system. The scope for its future enhancement rests on the shoulders of its creators. Our work in this field will continue to grow and we have a roadmap for adding more features to this classifier. Some of our planned enhancements are:

- Subtitle and transcript generation
- Changing video speed based on the action going on
- Vocal narration for disabled
- Large Curations and Sorting of videos
- Medical Video Analysis.

Acknowledgment

First and foremost, we would like to thank our Institute, Gokaraju Rangaraju Institute of Engineering and Technology, for providing us this opportunity and the environment to conduct the research for this paper. We would additionally like to thank the faculty members and peers of the college for their valuable feedback and assistance during our research process. Finally, we thank the UCF group, HMDB and Hollywood dataset creators for compiling the datasets and making them freely available to the public.

References

[1] https://karpathy.github.io/2015/05/21/rnn-effectiveness/

[2] http://www.image-net.org/

[3] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, arXiv:1704.04861 [cs.CV]

[4] Oord, Aaron van den; Dieleman, Sander; Zen, Heiga; Simonyan, Karen; Vinyals, Oriol; Graves, Alex; Kalchbrenner, Nal; Senior, Andrew; Kavukcuoglu, Koray (2016-09-12). "WaveNet: A Generative Model for Raw Audio". 1609. arXiv:1609.03499

[5] Zhe Wang, Kingsley Kuan and Mathieu Ravaut: Truly Multi-modal YouTube-8M Video Classification with Video, Audio, and Text, arXiv:1706.05461

[6] Emma An, Anqi Ji and Edward Ng : Large scale video classification using both visual and audio features on YouTube-8M dataset, unpublished.

[7] Ali Diba*, Mohsen Fayyaz*, Vivek Sharma, Amir Hossein Karami, Mohammad Mahdi Arzani, Rahman Yousefzadeh, Luc Van Gool : Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification, arXiv:1711.08200

[8] Bashivan, et al. "Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks." International conference on learning representations (2016).

[9] https://research.google.com/youtube8m/

[10] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A Large Video Database for Human Motion Recognition. ICCV, 2011.

[11] Khurram Soomro, Amir Roshan Zamir and Mubarak Shah, UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild, CRCV-TR-12-01, November, 2012.

[12] Marcin Marsza{\l}ek and Ivan Laptev and Cordelia Schmid : Actions in Context, IEEE Conference on Computer Vision \& Pattern Recognition, 2009

[13] http://publicdomainmovie.net/

[14] http://publicdomainflix.com/

[15] https://opencv.org/

[16] https://ffmpeg.org/about.html

[17] https://matplotlib.org/

[18] Karen Simonyan∗ & Andrew Zisserman+ Visual Geometry Group, Department of Engineering Science, University of Oxford {karen,az}@robots.ox.ac.uk, arXiv:1409.1556v6 [cs.CV] 10 Apr 2015

[19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna, Rethinking the Inception Architecture for Computer Vision, arXiv:1512.00567 [cs.CV]

[20] http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[21] https://machinelearningmastery.com/cnn-long-short-term-memory-networks/

[22] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, arXiv:1506.01497 [cs.CV].

[23] Hyeonwoo Noh, Seunghoon Hong, Bohyung Han. “Learning Deconvolution Network for Semantic Segmentation”, arXiv:1505.04366 [cs.CV]

[24] Shuiwang Ji ; Wei Xu ; Ming Yang ; Kai Yu, “3D Convolutional Neural Networks for Human Action Recognition”, IEEE Explore.

Mobility Management Challenges and Solutions in Mobile Cloud Computing System for Next Generation Networks

L. Pallavi[1], B. Thirumala Rao[3]
Department of CSE
Koneru Lakshmaiah Education Foundation
Vaddeswaram, India

A. Jagan[2]
Department of CSE
B V Raju Institute of Technology
Narsapur, India

***Abstract*—As of late, there is a dynamic improvement in the field of mobile computing, and mobile cloud computing (MCC) has been familiar with a potential development for portable administrations. Likewise, the mobile phones and their applications have high framework in the administration at any point and grow more rapidly. Again, MCC is depended upon to deliver on a very basic level increasingly inventive with multi applications. Moreover, Mobile handling incorporates versatile equipments, portable correspondence, cell programming, and right now there are various compact cloud applications. Versatility the board for supporting the area following and area based assistance is a significant issue of savvy city by giving the way to the smooth transportation of individuals and products. The mobility is valuable to contribute the development in both open and private transportation frameworks for keen urban communities. As the information is distributed computing and getting to it with cell phones all the exchanges experience the system so it is powerless against assault. For keeping the utilization of this fundamental apparatus of steady in this development world we are giving a portion of the answers for these difficulties to address in the field of MCC. In this paper, the main challenges faced by the enterprises and their corresponding solutions are discussed with the mobile cloud computing.**

Keywords*—Mobility management; energy consumption; network resource management; traffic management; security management

I. Introduction

Throughout late years, advances in the field of framework set up together handling and applications as for demand have incited a risky improvement of usage models, for instance, appropriated figuring, programming as an organization, arrange sort out, web store, and so forth [1]. As a vital application show in the hour of the Internet, Cloud Computing has transformed into a basic research subject of the intelligent and mechanical systems since 2007. Normally, appropriated processing is depicted as an extent of the organizations which are given by an Internet- based pack system. Such pack structures include a social occasion of negligible exertion servers or personal computers (PCs) [2], shifting through the various resources of the PCs as showed by a particular organization system, and offering ensured, strong, brisk, accommodating and direct organizations. For instance, data accumulating, getting to and figuring to clients [3]. In this time, PDAs are considered as the representatives of the distinctive mobile phones as they have been related with the Internet with the rapid creation of remote framework development. Inescapability and flexibility are two important features in the front line arrangement which gives an extent of the modified framework benefits through different framework terminals and strategies for getting to [4]. The inside development of disseminated processing is concentrating figuring, organizations, and express packages as an application to be offered like water, gas or capability to customers. Consequently, the mixture of a ubiquities adaptable framework and conveyed processing produces another enlisting mode, to be precise mobile cloud computing (mcc) [5]-[8]. In the meantime, the diverse programs reliant on bendy disseminated registering had been made and served to clients, as an example, Google Gmail, maps and navigation frameworks for cell, voice search, and some applications on an android degree, mobileme from Apple, live mesh from Microsoft, and Motoblur from Motorola.

The expanding situation towards MCC with the blast of versatile applications and the help of CC for various assortment of administrations for portable clients. MCC is presented as a combination of distributed computing with the portable figuring and cell phones. Be that as it may, alongside the helpfulness of this subject of portable distributed computing still research should be done on a few issues just as conceivable structures to help distributed computing on cell phones. MCC methods endeavor to concentrate on easing assets impediments in leaving innovation by utilizing distinctive systems of growth are screen increase, vitality enlargement, stockpiling expansion and application preparing. There are number of methodologies and contend that MCC handles that are expected to be the top of the line equipment, diminishes possession and upkeep cost, and lightens information wellbeing and client security of MCC [9]-[13]. The general model of MCC is given in Fig. 1.

In MCC, the past wireless based genuine enlisting, data storing and mass information taking care of have been traded to 'cloud' and subsequently the essentials of mobile phones in handling capacity and resources have been diminished, so the making, running, sending and using strategy for compact applications have been totally changed [14]. Of course, the terminals which people used to get to and get cloud organizations are suitable for PDAs like mobile phone, PDA, Tablet, and iPad yet not bound to settled contraptions, (for instance, PC), which reflects the ideal conditions and special objective of circulated figuring. Thus, from the two pieces of compact enrolling and circulated figuring, the adaptable conveyed processing is a mix of the two advances, an improvement of scattered, organized and thought estimations, and have extensive possibilities for application [15].

As shown is the Fig. 1, MCC can be basically isolated into distributed computing and portable processing. As the figuring and actual facts making ready tiers have been relocated to 'cloud', the ability necessity of cell phones is limited, some minimal effort cell phones or even non-advanced cellular phones can likewise accomplish transportable allotted computing by utilizing a cross-level mid-product [16]. Notwithstanding the truth that the patron in mcc is modified from desktops or settled machines to cellular phones, the fundamental concept remains dispensed computing. Flexible customers ship administration solicitations to the cloud via an internet browser or work location utility, the administration a part of cloud dispenses assets to the call for installation affiliation, while observing and figuring the factors of mcc might be completed to guarantee the qos till the affiliation is completed [17][18]. In MCC scene, an amalgamation of versatile processing and distributed computing correspondence systems makes a few complex difficulties. Albeit a portion of the difficulties, for example, consistent network, seller lock-in, versatility of the executives, security and protection are regular with portable registering and cloud computing [19]-[22]. A handoff the executives is required when a client moves starting with one remote cell then onto the next. At the point when a handoff happens inside the area of a homogeneous remote access innovation happening occasion is known as flat handoff and when this occasion happens among heterogeneous remote access arrange advances is known vertical handoff. Level handoff happens when the MNs are moving a long way from purpose of connection and go into the low flag quality territory in a homogeneous remote system. In a heterogeneous environment [23]-[25] (shown in Fig. 2), clients have a chance to get to the distinctive advances systems. A client might be profited by various system qualities which are not similar specifically. MN portability is upheld by vertical handoff as the correspondence innovation and access supporting foundation change. At some point vertical handoff happens because of client's comfort as opposed to inaccessibility of association. The handoff procedure turns out to be progressively intricate in such a domain contrasted with the homogeneous one.

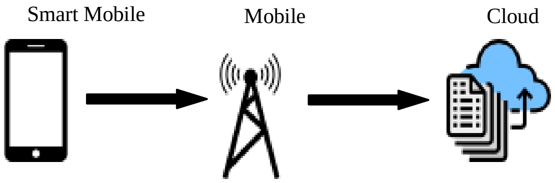


Fig. 1. Generic model of MCC.

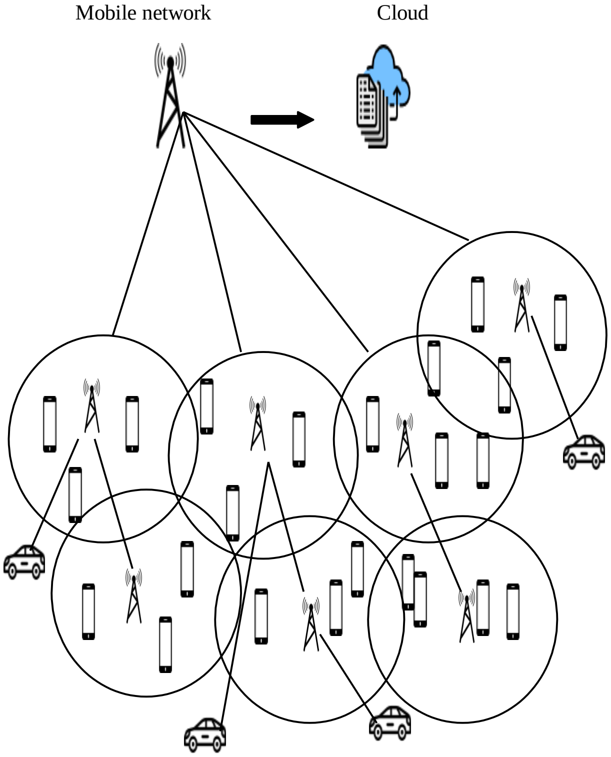


Fig. 2. MCC over Heterogeneous Wireless Data Networks.

This paper essentially spotlights on Mobility Management Challenges and Solutions in Mobile Cloud Computing (MCC) and a wide scope of audits work and endeavors on difficulties and arrangements in MCC are given. Likewise, this work features the principle illuminated research moves identified with cell phones, by taking certain difficulties and parameters like parcel misfortune, delay, correspondence cost, through put and system load. This work intends to be a valuable manual for MCC difficulties and issues and furthermore a point of reference to prepare for additional work and endeavors in MCC space. In second area, a foundation audit and outline is given. In third segment, issue proclamation and proposed philosophy is given. Ultimately, we finish up the paper.

II. Related Works

In this section, we review the year wise related works on mobility management challenges and solutions in MCC.

A. *Review between 2000 and 2005*

Corte et al. [26] have concentrated more on the administration of the QoS by utilizing the versatile programming specialists. They have given a reference philosophy which gives the systems administration condition where organize segments and the administrator can act with one another for the better nature of the work. The MAP framework which is pre-accessible in the portable programming stage have two distinct applications, one is to concentrate on asset reservation by utilizing RSVP and next is to give QoS to the pre-accessible traffic which is streaming in a virtual system. In view of this open system, further developed system can be utilized in an entirely adaptable manner.

However, the hindrance in this is it won't give the security ensure which should be given by the operator execution condition.

Manner et al. [27] have taken diverse procedure and conventions that has given QoS. They have investigated IntServ, DiffServ, blends of the first two things, RTP, INSIGNIA and ITSUMO. Every technique has diverse advantages and downsides. The third-age frameworks plan to give fast versatile access. These models are anyway shut frameworks and very constrained by government specialists and a couple of administrators. The BRAIN and MIND ventures, are likewise contemplating open IP-based portable remote system design. The tasks propose and assess an engineering that offers help for portable QoS, a structure for versatile applications, and a very much characterized interface through which applications can ask for the dimension of administration they require. The design has very much characterized interfaces to permit between operability among systems. Toward the finish of every technique portrayed definitely with the end that none of them truly gives both adaptable and exact QoS.

Hadjadj et al. [28] have grown an enthusiasm for the third era remote IP system and administration advances. This article gives open issues on the smooth and productive help of multicast sight and sound administrations including heterogeneous versatile IPv4 and IPv6 has. In the wake of investigating related takes a shot at this wide territory just as existing yet restricted and halfway arrangements, we have proposed and assessed an incorporated Multicast media passage for IPv4/IPv6 Multimedia Service Transition. The general M3G's design and related useful modules have been depicted and actualized on a Linux-based framework for highlights approval and execution assessment. Progress issues in regards to the client and the flagging control designs have been both tended to. As to the portability part of such administrations, new methodologies and enhancements have been added to the flagging arrangement through a superior coordination of existing sight and sound session the board, versatility and multicast flagging conventions. Thus, straightforward foundation and productive control of half and half interactive media multicast interchanges between portable IPv4-just and IPv6-just mists are conceivable while protecting inheritance IPv4 systems and administrations amid a smooth IPv6transition.With the remote, it required more data transfer capacity which it couldn't ready to give.

Mun et al. [29] have proposed versatile area the executives conspire called NHRP for an IP client on a move over wireless ATM organize. It depends on the NHRP standard, yet improved to help IP portability in wireless ATM systems which have qualities that the purpose of connection to ATM systems changes as often as possible after some time. Rather than the NHRP, where just thinks about settled customers, the recovered location data about the versatile customer from the location goals server in the proposed plan mirrors the present area of the portable customer. Accordingly, the association can be set up straightforwardly in ATM layer. The blame tolerant steering capacity of the proposed plan by joining with either portable PNNI or LRs. Investigative models are determined to affirm the execution of the proposed plan. The execution is estimated as far as the expense. Hunt cost and following expense are gotten. This plan is appeared to be viable than the novel blends. This plan is additionally acquired by the variety of ATM and IPOA scope parameters S and D.

Vincent Lenders et al. [30] have given a novel methodology towards productive and strong administration revelation in portable specially appointed systems. This plan isn't an augmentation to existing impromptu steering conventions or an adaption of leaving administration revelation systems, yet presents and assesses new ideas. All things considered, electric field-based administration revelation utilizes a basic instrument to locate the best course to the nearest administration example: at every hub, the demand is steered towards the steepest angle until it achieves the administration occurrence. The significant favorable position of this methodology be that as it may, is its straightforwardness and lucidity in structure. This technique to perform administration disclosure can without much of a stretch be adjusted to be utilized for extra errands. The main application that rings a bell is parcel directing in MANETs. Rather than sending solicitations to support examples just, a similar component can be utilized to build up correspondence between two gadgets as long as they are exceptionally recognized in the system. Another important property of this methodology is its freedom of the basic system convention. In fact, it isn’t just autonomous; it works even without any basic directing convention. Table I shows the summary of reviews between 2000 and 2005.

TABLE. I. SUMMARY OF REVIEWS BETWEEN 2000-2005

Ref. No.	Year	Algorithm	Drawback	Performance
26	2001	Mobile agents	Security guarantee	Flexible way to access
27	2002	Access Networks, MHs,	Didn’t provide QoS	Fast adoption, fast mobility
28	2003	Multi-media gateway	lesser bandwidth	Wireless network
29	2004	wireless ATM technology	with higher data, the network is not stable	Fast speed for wireless
30	2005	Mobile and hoc networks	low load balancing	network is stable, even when conditions are dynamic

B. *Review between 2006 and 2008*

The requirement of ubiquitous internet get admission to (e.g., in public transportation systems) is growing. Consequently, mechanisms that permit complete IP networks to be cellular without breaking ongoing connections of the nodes of the internet- work are wanted. The IETF NEMO WG has come up with an IP-level network mobility answer: The Network Mobility (NEMO) [31]. Fundamental Support convention that empowers a network to change its purpose of connection. In this paper we have built up a usage of the NEMO Basic Support convention for Linux, and we have utilized that to tentatively assess the execution of the convention. The NEMO Basic Support convention fundamentally comprises in setting a bidirectional passage between the MR and it's HA. This passage adds both starts to finish postponement and bundle overhead. This postponement can be unsatisfactory for some ongoing applications, yet in addition influences the general TCP execution, as it has been for all intents and purposes appeared in this paper. Moreover, the additional parcel overhead expands the data transfer capacity prerequisites for applications. For instance, run of the mill 64 kbps connections would not have the capacity to deal with VoIP Skype calls of a hub having a place with a two-level settled system. Be that as it may, in overlay systems, choice of top notch essential and reinforcement ways is a testing issue because of the sharing of physical connections among overlay ways. Such sharing influences many steering measurements, including joint disappointment likelihood and connection blockage. Henceforth to defeat this issue taken care of by administration Clouds engineering with Planet Lab Internet proving ground alone and a versatile figuring testbed. This paper depicts Service Clouds, a disseminated framework intended to encourage fast prototyping and organization of versatile correspondence administrations. The foundation consolidates versatile middleware usefulness with an overlay organize substrate so as to help dynamic instantiation and reconfiguration of administrations. The Service Clouds design incorporates an accumulation of low-level offices that can be summoned straightforwardly by applications or used to make progressively complex administrations [32].

The architecture of a SDCR system's [33] highlights and CWC were advertised. At that point, as an attainability investigation of SDCR terminal, this paper additionally presented the arrangement and elements of the HWP, SWP, and waveforms and tended to a deliberate information for range detecting period and reconfiguration period by utilizing genuine waveforms. The SDCR terminal executed detecting and reconfiguration of radio correspondence frameworks over 400MHz-6GHz. The normal reconfiguration time for every correspondence framework is at most 1650ms and real detecting time is inside a few seconds. This improvement is the world's first work. As a further work, field test by utilizing the prototypeon the CWC must be required. These internetworking models might be ordered as tight coupling, free coupling, and distributed systems administration (additionally alluded as no coupling). In any case, these methodologies appear to give constrained internetworking capacity as neither of these structures has effectively tended to the issue of consistent continuation of administrations. So, they presented an [34] internetworking model for WLAN and 3G cell systems with the 3GPP's IMS structure going about as a referee. It tended to numerous insufficiencies of the current internetworking designs. The most noteworthy advantage is its capacity for arranging and overseeing ongoing sessions with the utilization of the IMS as a brought together session controller. IMS-SIP based terminal and session portability was inspected inside the extent of this system for two situations; that is, while meandering from UMTS to WLAN and from WLAN to UMTS. Results got from the OPNET based recreation stage showed the conduct of handoff for these situations. Reenactment results showed that a make-before-break type handoff from UMTS to WLAN is equipped for giving worthy dimensions of consistent congruity of administrations. Results likewise demonstrated a circumstance with information duplication, which must be tended to at a lower layer. Notwithstanding, the break-before-make type handoff situation from WLAN to UMTS demonstrated a short interim of administration intrusion because of its non-covered nature of inclusion.

These days versatile correspondences acquire and more significance. The expanded use requires increasingly complex administrations. In this paper conceivable, future system design is inspected, called BIONETS [35]. For this situation the versatile hubs for the whole system, without devoted spine are available. They seek after the objective of finding an ideal data spread model over some versatility demonstration. The examination covers recently characterized convention called IOBIO, established communicate and a recently created versatile communicate calculation. We run a few recreations - with a claim test system made in OMNeT++ so as to choose which one is the ideal data spread strategy for the given versatility condition. The outcomes give us the favorable position to additionally enhance the correspondence in such systems. Notwithstanding, remote handheld gadgets with compelled usefulness, for example, little screen, constrained figuring power, etc. limit the client arranged QoS accommodated remote web surfing. It is troublesome for remote portable clients to peruse vast website page intended for PC clients easily. Consequently by alluding to tremendous processing capacity and capacity asset of distributed computing foundation, another remote web get to mode is proposed [36]. Right off the bat, the framework structure is available. In this manner, the two key segments of framework are portrayed in detail: the one is disseminated website page adjustment motor, which is intended for the reason that the motor can be conveyed by registering cloud dispersed and parallel; the other is circulated site page squares the executives dependent on distributed computing, which is proposed so the site page adjustment motor can be sent sensibly. Besides, a model framework and a lot of assessment tests have been actualized.

Portability has acquainted another measurement with the remote research zones, for example, IP versatility the board conventions and remote specially appointed systems. In this paper they investigate the issues identified with the portability in wireless sensor networks (WSNs). The point of this work is to distinguish upgrades that can be acquired thinking about versatility, perceive its examination challenges. We depict distinctive dimensions of versatility in WSNs and feature the impact of the portability on the execution of WSN conventions [37]. Quality of Service (QoS) provisioning in remote systems rapidly prompts versatility issue because of hard asset portion for every session. Heterogeneous remote condition gives elective assets, however the multifaceted nature of taking care of such framework is high. Henceforth they exhibited a virtual resonance provisioning (VRP) conspire that works together with gathered remote systems in regards to asset accessibility and evaluations asset discharge in not so distant future to keep up prescient asset list for ideal asset usage. Asset usage is additionally upheld by fine-grained QoS administration sub-classes, plan to make exact asset allotment. These sub-classes structure subset of QoS classes and expend assets on exact requirements of administration and arrangements of physical connections. Results demonstrate that a lot of higher call affirmation rates are accomplished by incorporating virtual asset provisioning and load-adjusting approach [38].

Opportunistic networks, in which hubs shrewdly abuse any pair-wise contact to recognize next bounces towards the goal, are a standout amongst the most intriguing innovations to help the unavoidable systems administration vision. Opportunistic networks permit content sharing between versatile clients without requiring any prior Internet foundation, and endure allotments, long disengagements, and topology unsteadiness as a rule. In this paper they proposed [39] a setting mindful structure for directing and sending in shrewd systems. The structure is general, and ready to have different kinds of setting mindful directing. In this work they additionally present a specific convention, HiBOp, which, by abusing the system, learns and speaks to through setting data, the clients' conduct and their social relations, and utilizations this information to drive the sending procedure. The examination of HiBOp with reference to elective arrangements demonstrates that a setting mindful methodology dependent on clients' social relations ends up being a proficient answer for sending in shrewd systems. We show execution upgrades over the reference arrangements both as far as asset usage and as far as client saw QoS. The methodology appears to be encouraging, in spite of the fact that it is hard to assess its execution with exceedingly portable hubs since it is hard to make sure of n the exactness of the two-bounce neighborhood of a hub. Consequently, to conquer this issue they displayed a far-reaching overview of the best in class for vehicle impromptu systems [40]. They begin by evaluating the conceivable applications that can be utilized in VANETs, to be specific, wellbeing and client applications, and by recognizing their prerequisites. At that point, they characterize the arrangements proposed in the writing as indicated by their area in the open framework interconnection reference show and their relationship to wellbeing or client applications. They break down their focal points and deficiencies and give proposals to a superior methodology. They likewise portray the diverse techniques used to recreate and assess the proposed arrangements. At long last, they finish up with recommendations for a general design that can shape the reason for a commonsense VANET. Table II shows the summary of reviews between 2006 and 2008.

C. Review between 2008- 2010

Here searched about notions such as interlocking directorships, communities of practice, learning regions and labor mobility [41]. Here checked on ebb and flow inquire about on information the board and learning move with regards to developments. Explicit consideration is centered around the joining of the board viewpoints into the travel industry look into. They investigated a portion of the key instruments and courses of learning exchange inside the travel industry. There is likewise a developing examination plan on learning the board inside the travel industry however advance is variable with most research being inside the inn segment, where a scope of ongoing investigations have analyzed parts of information exchange. They additionally attract regard for the need to give nearer thoughtfulness regarding the idea of advancements inside the travel industry and to consider these in a learning the executives system. Versatility the board issue happens consequently to beat this issue a practical reenactment of portability for urban remote systems is tended to [42]. As opposed to most other portability displaying endeavors, the majority of the parts of the exhibited versatility model and model parameters are gotten from overviews from urban arranging and traffic designing exploration. The portability demonstrates talked about here is a piece of the UDel Models, a suite of devices for sensible reproduction of urban remote systems. The UDel Models reenactment instruments are accessible on the web.

TABLE. II. SUMMARY OF REVIEWS BETWEEN 2006 AND 2008

Ref. No.	Year	Algorithm	Drawback	Performance
31	2006	IETF in NEMO	Node mobility, delay	Improved NEMO, Reduced delay
32	2007	Service clouds	Joint failure chance and hyperlink congestion	Packet loss rate discount, robustness stepped forward overhead delay reduced.
33	2007	software defined cognitive radio (SDCR)	Drawback of the distance and it is also difficult to consist of spectrum sensing module	Better feasibility, Increased space
34	2007	WLANs by using the IP Multimedia Subsystem (IMS)	Limited internetworking capability	Better performance, improved capability
35	2007	BIONETS/ IOBIO algorithm	Limited number of mobility pattern	High efficiency
36	2008	wireless web access mode	Wi-fi hand-held devices with restricted functionality such as small display, limited computing power	huge computing ability and storage resource of cloud computing infrastructure
37	2008	routing protocol proposed for sensor networks	buffering time	Buffering time is high, message delay reduced
38	2008	Virtual resource provisioning (VRP)	Heterogeneous wireless environment complexity	Flexible mobility, reduced complexity
39	2008	Context-conscious framework for routing and forwarding in opportunistic networks	Stability, single point failure, scalability	High stability ,High networking overhead
40	2008	VANETs	Robust transmission, packet collision	Reduced collision, Stability improved

A net-centric announcement, command, and control architecture for a heterogeneous unmanned aircraft system comprised of small and miniature unmanned aircraft are given [43]. An incorporated framework was created utilizing a base up configuration way to deal with reflect and improve the interchange organized correspondence and self-sufficient airplane coordination. The upsides of the methodology are exhibited through a portrayal of the unmanned framework that came about because of utilizing this structure procedure. The equipment framework is portrayed, including both little and smaller than usual unmanned flying machine alongside their particular aeronautics' frameworks. System engineering is depicted that flawlessly consolidates the little airplane's IEEE

802.15.4 components with IEEE 802.11 (WiFi) segments on the little flying machine. Reconciliation of intra vehicle correspondence and administration disclosure is likewise depicted. Various leveled control engineering is exhibited that utilizes the system design to organize the little and small-scale flying machine at a few layers in the control chain of command. Equipment on the up and up shows are performed to approve the abilities of the heterogeneous unmanned airplane framework. They abuse the particular data accessible by the MCC [44] the clients' area, setting, and asked for administrations, and altogether advance the Heterogeneous Access Management plans created for the customary heterogeneous access situations.

They presented design for a portal to provide remote management access to virtualized device management servers hosted in a service cloud [45]. This plan is focused to conceal the subtleties of the gadget the executives behind a standard-based, uniform control interface that can be a cross-stage specialist that can keep running on numerous versatile stages. They will likewise portray quickly the model they are creating as an inner pilot. The Mobile DM programming and the end clients caused a great deal of issues and sobs for critical enhancements. The engineering of SCM [46] a mix of two developing they advances, semantic and distributed computing, for upsetting information access and handling capacities over portable stage. The accentuation lies on decoupling of information preparing and the board from versatile equipment, alongside the way SCM can be utilized to determine issues that have continued over years, and investigating new open doors that SCM may guarantee to offer issues uninformed quality DM programming.

In a Hadoop based system for impromptu versatile distributed computing – they explicitly allude to an exploration paper and the creators' choice to utilize and port Hadoop to construct a virtual distributed computing supplier for cell phones [47]. First and foremost, they are going to layout a short presentation and list the difficulties that they face actualizing a structure for portable distributed computing. They finish up with a short case of other Map-Reduce based MCC system, which accomplished better execution utilizing its very own custom usage. The serious issue is the manner by which to change Hadoop system. To defeat its issue has high data quality. The issue is the extensive scale they present GenLM, a permit the board arrangement reasonable for these situations [48]. It has been worked so as to give a protected and powerful answer for ISVs that need to stretch out their product use to these frameworks. They give ISVs an apparatus chain to actualize self-assertive programming permitting models. In the meantime, they guarantee that licenses are versatile, for example they can be utilized on any asset the client approaches. The issue is the substantial scale.

They overlook the (IMS) as a common coupling mediator for real-time consultation negotiation and management [49]. Queuing idea is based totally on analytical model for assessing the execution of vertical handoff the executives Bett Heyen has interworked 0. 33 technology (3G) mobile systems and the Wlans is to be had. The investigation includes vertical handoff execution estimates, as an instance, delay, temporary parcel misfortune, jitter, and flagging overhead/fee. The closing piece of this paper introduces some effects from opnet based recreations to take a look at the logical model and effects. The progressed session control is to propel the making of client subgroups, subsets of a comparable substance collect depending on machine, client and condition putting to the diploma that these are critical for effectiveness [50]. Their traits inside the form of content descriptions are communicated to the media transport to achieve the suitable content material for every session. Table III shows the summary of reviews between 2008 and 2010.

D. Review between 2011 and 2014

Li et al. [51] proposes Urban-traffic the executive's framework utilizing savvy traffic mists to control the traffic in the urban territories. Operator based traffic cloud the board frameworks can utilize the self-rule, portability, and flexibility of versatile specialists to manage dynamic traffic situations. Distributed computing can enable such frameworks to adapt to a lot of capacity and figuring assets required to utilize traffic cloud technique specialists and mass transport information viably. This article audits the historical backdrop of the improvement of traffic cloud control and the board frameworks inside the advancing processing worldview and demonstrates the condition of traffic cloud control and the executive's frameworks dependent on versatile multi-operator innovation. Astute transportation mists could give administrations, for example, choice help, a standard advancement condition for traffic cloud the executives methodologies, etc. With portable operator innovation, an urban-traffic the executives' framework dependent on Agent-Based Distributed and Adaptive Platforms for Transportation Systems (Adapts) is both achievable and adequate. Be that the substantial scale utilization of portable operators will prompt the rise of a mind boggling, amazing association layer that requires gigantic registering and power assets. To manage this issue, we propose a model urban-traffic the board framework utilizing savvy traffic cloud mists.

TABLE. III. SUMMARY OF REVIEWS BETWEEN 2008 AND 2010

Ref. No.	**Year**	**Algorithm**	**Drawback**	**Performance**
51	2009	Innovation with tourism	High cost	Knowledge management framework
52	2009	Urban wireless network technique	Slow network	Network communication and autonomous aircraft communication
53	2009	Heterogeneous aircraft technique	Miniature aircraft is very small	Capability of heterogeneous unnamed aircraft system
54	2010	Mobile cloudcontroller	It has very low quality	Intelligent radio access management
55	2010	Virtualized device management technique	It cause the end users	Remote management
56	2010	Theyb technologies	Low information quality	Data processing and management form mobile hardware
57	2010	Virtual cloud computing technique	The problem is to how to change the hadoop framework	Custom management
58	2010	Hadoop framework	It has very large-scale	License management
59	2010	Cellular network and WLAN Technology	It has been delay in information	Vertical handoff management between interwork generation
60	2010	Context-awaremulticast session	Low efficiency	Session management

Sardis et al. [52] have acquainted a novel procedure with enhance the QoS. Ongoing advances in cell phones and system innovations have set new patterns in the manner in which we use PCs and access systems. Cloud computing, where handling and capacity assets are living on the system is one of these patterns. The other is Mobile Computing, where cell phones, for example, advanced mobile phones and tablets are accepted to supplant PCs by joining system availability, versatility, and programming usefulness. Later on, these gadgets are relied upon to flawlessly switch between various system suppliers utilizing vertical handover instruments so as to keep up system availability consistently. This will empower cell phones to get to Cloud Services without interference as clients move around. Utilizing current administration conveyance models, cell phones moving starting with one topographical area then onto the next will continue getting to those administrations from the nearby Cloud of their past system, which may prompt moving a vast volume of information over the Internet spine over long separations. Cloud-Based Mobile Media Service Delivery in which administrations keep running on limited open Clouds and are fit for populating other open Clouds in various topographical areas relying upon administration requests and system status. Utilizing an investigative system, it is contended that as the interest for explicit administrations increments in an area, it may be progressively productive to draw those administrations nearer to that area. This will keep the Internet spine from encountering high traffic stacks because of sight and sound streams and will offer specialist organizations a mechanized asset portion and the executive's system for their administrations.

According to Amoroso et al. [53] in 90's, the endeavor systems were utilized to interface people in general Internet in huge numbers, firewalls shielded confided in substances inside the venture from untrusted access by outside elements. The methodology steadily advanced into the well-known venture security demonstrate, with parts, for example, intrusion detection systems (IDSs)/intrusion prevention systems (IPSs), antivirus (AV) and antispam (AS) channels, risk the board frameworks, and information spillage counteractive action (DLP) devices giving extra help. The subsequent border arranges demonstrate has been a mainstay of insurance structure for security draftsmen for almost three decades. Be that as it may, generally speaking venture trust in this setup has persistently corrupted because of network choices and advancing dangers. Here by analyzing a technique to reestablish this trust utilizing centered insurance methodologies.

Panta et al. [54] propose a framework that makes crafty utilization of portable registering gadgets and specially appointed systems administration to give a transient stockpiling administration to customers in a confined geological area. The primary test is to counterbalance the potential information misfortune brought about by hub versatility with between hub correspondences. The content based on reproduction and hypothesis, such as an administration is possible, is given an adequately high thickness of cell phones. A conveyed correspondence and capacity convention is then utilized for circumstances where every single cell phone are inside correspondence scope of one another, and it is appeared through proving ground investigations and recreation that the convention works accurately and makes effective utilization of storage room and correspondence transfer speed, while expanding the life span of put away information.

Gkatzikis et al. [55] depicted Contemporary cell phones produce substantial heaps of computationally serious undertakings, which can't be executed locally because of the restricted handling and vitality abilities of every gadget. Cloud offices empower cell phones customers to offload their assignments to remote cloud servers, bringing forth MCC. The test for the cloud is to limit the assignment execution and information exchange time to the client, whose area changes because of portability. The nature of administration ensures is especially testing in the dynamic MCC condition, because of the time-differing transfer speed of the entrance interfaces, the regularly changing accessible handling limit at every server and the time shifting information volume of each virtual machine. In this article, advocate the requirement for novel cloud models and relocation instruments that viably bring the registering intensity of the cloud nearer to the versatile client. This strategy considers a distributed computing design that comprises of a back-end cloud and a nearby cloud, which is joined to remote access framework (for example LTE base stations). It diagrams distinctive classes of assignment movement arrangements, traversing completely clumsy ones, in which every client or server self-sufficiently settles on its relocation choices, up to the cloud-wide relocation procedure of a cloud supplier. We finish up with a talk of open research issues in the region.

Taleb et al. [56] analyzed that colossal increment in portable information traffic, there is a general pattern toward the decentralization of versatile administrator systems, at any rate partially. This will be additionally encouraged with the virtualization of portable system capacities and the empowering of versatile cloud organizing, whereby versatile systems are made on interest and in an adaptable way. Versatile system decentralization won't be proficient without reconsidering portability the board plans, especially for clients moving for a long separation as well as at a rapid To help such exceedingly portable clients, this paper presents: 1) an information stay entryway (GW) migration strategy dependent on client versatility, history data, and client movement examples, and 2) a handover the executives approach that chooses an objective base station or advanced Node B (eNB) in an approach to limit versatility grapple GW movement. The execution of this plans is assessed by means of Markov show based investigation and through reproductions. Empowering results are gotten, approving the plan targets of the plan.

Zhang et al. [57] depicted telecom cloud give more consideration on area-based applications and administrations. Because of the irregularity and fluffiness of human portability, despite everything it stays open to anticipate client versatility. In this article, it examines the substantial scale client portability follows that are gathered by a telecom administrator. It is discovered that versatile call designs are profoundly associated with the co-area designs at a similar cell tower in the meantime. By extricating such social associations from cell call records put away in the telecom cloud, and further propose a portability expectation framework that can keep running as a foundation level administration in telecom cloud stages. They lead several contextual analyses on portability mindful personalization and prescient asset assignment to expound how this framework drives another method of versatile cloud application.

According to Qi et al. [58] cell phones are turning into the essential stages for some clients who dependably meander around while getting to the distributed computing administrations. From this, the distributed computing is incorporated into the versatile condition by presenting another worldview, portable distributed computing. With regards to portable figuring, the battery life of cell phone is constrained, and it is essential to adjust the portability execution and vitality utilization. Luckily, cloud administrations give the two chances and difficulties to versatility the board. Taking the exercises of cloud administrations getting to into thought, the creators utilized an administration mindful area refresh instrument, which can distinguish the nearness and area of the cell phone without conventional occasional enrollment refresh. Expository model and reproduction are created to examine the new system. The outcomes show that the administration mindful area refresh the board can decrease the area refresh times and handoff flagging, which can effectively spare power utilization for cell phones.

Che et al. [59] looks at that portability is nature on the planet and has advanced into a characteristic element and a key main impetus of things to come system, albeit possibly treated as one viewpoint in a particular system when it started in a cell framework. Going up against developing correspondence ideal models, for example, versatile informal organizations, portable distributed computing, Internet-of-Things, and the desire for universal and consistent network, the current portability the executives advancements face issues, for example, work repetition, framework unpredictability, and wastefulness. The plan reasoning of vertical decoupling the portability substance into administration element and gadget element, and even decoupling the element personality and locator-identifier, are both examined. The plan theory is connected in this capacity reference model and convention reference model of MDN to extract the portability bolster capacities, elements, and standards. At last, the open issues in MDN are talked about.

The Internet of Things (IoT) is rising as one of the real patterns for the following advancement of the Internet, where billions of physical items or things (counting yet not constrained to people) will be associated over the Internet, and a tremendous measure of data information will be shared among them (Yue et al. [60]). Be that as it may, the present Internet was based on a host-driven correspondence show, which was principally intended for fulfilling the need of pair-wise distributed interchanges and can't well oblige different propelled information driven administrations supported by the IoT in which clients care about substance and are unmindful of areas where the substance is put away. Web dependent on information-centric networking (ICN) is called DataClouds, to more readily suit information driven administrations. Not the same as existing ICN-based structures, by taking the sharing idea of information driven administrations under the IoT into thought and present consistently and physically framed networks as the fundamental building squares to develop the system with the goal that information could be all the more proficiently shared and spread among intrigued clients. It is additionally expand on a few basic structure difficulties for the Internet under this new design and demonstrate that DataClouds could offer more productive and adaptable arrangements than conventional ICN-based models. Table IV shows the summary of reviews between 2011 and 2014.

TABLE. IV. SUMMARY OF REVIEWS BETWEEN 2011 AND 2014

Ref. No.	Year	Algorithm	Drawback	Performance
51	2011	City-site visitors c control system using clever visitors clouds	Cost is high, operating frequency is up to 2.66-ghz, several pcs or a high performance server are needed to deal with the experimental scale of several hundreds of intersections	High performances between the number of intersections and evolution time
52	2013	Cloud-based cell media provider shipping	now not scale to cowl the future desires of cell customers, lack of latency as a user movements while streaming a video	efficient management of community resources while offering a excessive QoS for the clients
53	2013	Focused protection strategies	Cost implications of this technique is high, complexity will at once relate to the diploma of use.	The trust degree is high as compared to the other technique.
54	2013	Distributed communication and storage protocol	Offset the ability information loss resulting from node mobility	Analyzed the parameters like effect of node mobility and failures, convergence time, effect of limited node capacity,
55	2013	Cloud computing structure that includes a returned-stop cloud and a nearby cloud	Decrease the undertaking execution and records switch time to the consumer, large-scale datacenter networks which include multiple dispersed server centers.	Efficient task scheduling
56	2014	Data anchor gateway (GW) relocation method, a handover management policy	Data traffic and the interference	Mobility, selection of target base station
57	2014	Mobility prediction machine that may run as an infrastructure-degree carrier	Difficult to identify the similar patterns when they are too congested.	The massive-scale user mobility lines which can be amassed by using a telecom operator, cellular call styles.
58	2014	Service-aware location update mechanism	Battery life of mobile device is restricted, mobility performance is also limited	Service-aware location update management is performed.
59	2014	Mobility-driven network (MDN)	Function redundancy, system complexity, and inefficiency	Analysis of the mobility support functions
60	2014	Architecture based on information-centric networking (ICN	Pair-wise peer-to- peer communications are implemented in advance	For the better performances of data-centric services.

E. Review between 2015 and 2018

Li et al. [61] to more readily suit information driven administrations. Not the same as existing ICN-based structures, by taking the sharing idea of information driven administrations under the IoT into thought and present consistently and physically framed networks as the fundamental building squares to develop the system with the goal that information could be all the more proficiently shared and spread among intrigued clients. It is additionally expand on a few basic structure difficulties for the Internet under this new design and demonstrate that DataClouds could offer more productive and adaptable arrangements than conventional ICN-based models.

Aissioui et al. [62] have taken a shot at 5G versatile system design is required to offer abilities to oblige the relentless ascent in portable information traffic and to meet further stringent inactivity and dependability necessities to help assorted high information rate applications and administrations. MCC in 5G has risen as a key worldview, promising to increase the capacity of cell phones through provisioning of computational assets on interest, and empowering asset compelled cell phones to offload their preparing and capacity prerequisites to the cloud framework. Follow Me Cloud (FMC), thusly, has risen as an idea that enables consistent relocation of administrations as per the comparing clients' versatility. In the interim, Software Defined Networking (SDN) is a new worldview that allows the decoupling of the control and information planes of customary systems and gives programmability and adaptability, enabling the system to progressively adjust to changing traffic examples and client requests. While the SDN usage is picking up force, the control plane is as yet experiencing adaptability and execution worries for an extremely expansive system. These adaptability and execution issues with regards to 5G versatile systems by presenting a novel SDN/OpenFlow-based design and control plane structure custom fitted for MCC-based frameworks and all the more explicitly for FMC-based frameworks where portable hubs and system administrations are liable to imperatives of developments and relocations. In spite of an incorporated methodology with a solitary SDN controller, this methodology allows the circulation of SDN/OpenFlow control plane on a two-level progressive engineering: a first level with a global controller G-FMCC, and second level with several local controllers LFMCC(s).

Rahimi et al. [63] have taken a shot at the ideal and reasonable administration distribution for an assortment of portable applications (single or gathering and communitarian versatile applications) in portable distributed computing. The structure to show portable applications as an area location-time workflows (LTW) of errands; here clients versatility designs are meant portable administration use designs. An ideal mapping of LTWs to layered cloud assets considering different QoS objectives such application delay, gadget control utilization and client cost/cost is a NP-difficult issue for both single and gathering based applications. A proficient heuristic calculation considered MuSIC that can perform well (73% of ideal, 30% superior to straightforward techniques), and scale well to a substantial number of clients while guaranteeing high versatile application QoS. At that point assess MuSIC and the 2-level versatile cloud approach by means of usage (on certifiable mists) and broad reenactments utilizing rich portable applications like serious flag preparing, video gushing and mixed media record sharing applications.

Kumar et al. [64] have worked with the across the board ubiquity and utilization of ICT around the globe, there is expanding enthusiasm for supplanting the conventional electric lattice by the brilliant network sooner rather than later. Many savvy gadgets exist in the shrewd matrix condition. These gadgets may impart their information to each other utilizing the ICT-based foundation. The examination of the information produced from different keen gadgets in the savvy lattice condition is a standout amongst the most difficult assignments to be executed as it changes as for parameters, for example, measure, volume, speed, and assortment. The yield of the information examination should be exchanged to the end clients utilizing different systems and keen machines. Be that as it may, at times systems may wind up over-burden amid such information transmissions to different savvy gadgets. Thus, huge deferrals might be caused, which influence the general execution of any actualized arrangement in this condition. We explore the utilization of VDTNs as one of the answers for information dispersal to different gadgets in the shrewd network condition utilizing portable edge registering. VDTNs utilize the store-and-convey forward component for message scattering to different keen gadgets so that postponements can be diminished amid over-burdening and clog circumstances in the center systems.

Ha et al. [65] have taken a shot at one of the significant difficulties in 6LoWPAN is to give nonstop administrations while versatile hubs' developments with limiting system blocked off time caused because of handoffs. Despite the fact that MIPv6, HMIPv6, and PMIPv6 are regularly acknowledged principles to address this in IP systems, they can't characteristically evade the corruption in correspondence quality amid handoff, since they are not planned with thought of compelled hub systems like 6LoWPAN. A quick versatility, the executives' convention for 6LoWPAN, named intra-MARIO is proposed. To limit handoff deferral and improve administration accessibility, intra-MARIO presents three critical parts, which are a quick rejoin plot for handoff the board with a versatile surveying-based development discovery and multi-bounce pointer sending plans for area the executives. To legitimize the adequacy, they have led the broad recreations by contrasting intra-MARIO and earlier plans like an essential versatility the executives plot and a PMIPv6-based convention.

Mazza et al. [66] have dealt with expanding urbanization dimension of the total populace has driven the advancement of a brilliant city geographic framework, imagined as a completely associated wide region portrayed by the nearness of a large number of shrewd gadgets, sensors, and handling hubs went for circulating knowledge into the city. In the meantime, the inescapability of remote advances has prompted the nearness of heterogeneous systems, working at the same time in a similar city region. The UMCC system is produced, presenting a portable distributed computing model portraying the flows of information and tasks occurring in the savvy city. A unified offloading component is utilized where correspondence and figuring assets are mutually overseen, permitting load adjusting among the different substances in the earth, appointing both correspondence and calculation undertakings so as to fulfill the savvy city application prerequisites.

Jeon et al. [67] have taken a shot at plan and sending of novel portable system engineering, inspired by the difficulties getting from the unstable increment in information traffic on administrator systems, is a squeezing issue in the present media communications. Distributed mobility management (DMM) acquaints a key thought with handle the traffic bottlenecks that sway current portable systems, by proposing the arrangement of circulated portability stay directs close toward terminal areas. A far-reaching examination and correlation consider methodically breaking down accessible plan alternatives have not yet been given. At that point distinguish fundamental structure contemplations and their hidden alternatives, contrasting their effect on client and system execution.

Ahmad et al. [68] have dealt with MCC is a creating innovation that helps with enhancing the nature of the versatile administrations. Since the expansion in portable assets, the specialists have stepped up with regards to take into consideration asset sharing among heterogeneous cell phones. Subsequently, to structure framework design for versatility models and asset sharing are key issues that require most extreme endeavors to be settled to accomplish foreseen destinations. Framework engineering dependent on the various leveled asset sharing component for MCC and the framework design is isolated into three spaces, for example, Global Cloud Server (GCS), Local ISP Server (LIS), and Gateway Server (GWS). A foglet selection scheme is evolved based totally on the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) decision mechanism. Diverse parameters consisting of delay, jitter, bit error rate, packet loss, communication cost, response time, and network load are considered for choosing the most appropriate network.

Chaudhary et al. [69] have taken a shot at, exponential increment in the quantity of Internet-empowered gadgets, which has brought about prominence of haze and distributed computing among end clients. End clients expect high information rates combined with secure information access for different applications executed either at the edge (haze processing) or in the center system (distributed computing). Be that as it may, the bidirectional information stream between the end clients and the gadgets situated at either the edge or center may cause blockage at the cloud server farms, which are utilized mostly for information stockpiling and information investigation. The high portability of gadgets (e.g., vehicles) may likewise represent extra difficulties as for information accessibility and handling at the center server farms. Henceforth, there is a need the majority of the assets accessible at the edge of the system to guarantee the smooth execution of end-client applications. The NSC administration demonstrate mechanizes the virtual assets by affixing in an arrangement for quick processing in both registering advances. This engineering additionally bolsters information investigation and the executives concerning gadget versatility. Besides, we additionally contrast the center and edge registering and admiration to the kind of hypervisors, virtualization, security, and hub heterogeneity. By concentrating on hubs' heterogeneity at the edge or center in the 5G condition, we additionally present security difficulties and conceivable kinds of assaults on the information shared between different gadgets in the 5G condition.

Enayet et al. [70] have chipped away at, the shrewd city idea, which includes numerous controls, for instance, brilliant human services, keen transportation, and savvy network, has turned out to be famous on account of its capacity to enhance urban residents' personal satisfaction. In any case, most administrations in these regions of shrewd urban areas have progressed toward becoming information driven, in this way producing huge information that require consistent ongoing access, sharing, putting away, preparing, and examination anyplace whenever for clever basic leadership to enhance expectations for everyday comforts. MCC can assume an imperative job by enabling a cell phone to get to and on-load huge information related errands to incredible cloudlet servers joined to numerous remote APs, along these lines guaranteeing that the QoS requests of end clients are met. The availability of cell phones with a given AP isn't consistent, yet rather sporadic with shifting sign qualities. Besides, the heterogeneity of the cloudlet assets and the enormous information application demands place extra difficulties in settling on ideal code execution choice. Table V shows the summary of reviews between 2015 and 2018.

TABLE. V. SUMMARY OF REVIEWS BETWEEN 2015 AND 2018

Ref. No.	Year	Algorithm	Drawback	Performance
61	2015	Mechanisms and demanding situations on Mobility-Augmented Service Provisioning for Mobile Cloud Computing	Pair-wise peer-to-peer communications are not implemented in advance.	Enhanced capacity of poor – resource mobile devices by using service provided in remote cloud
62	2015	FMC based on LISP (Local/Identifier Separation Protocol).	Difficult to identify the similar patterns when they are too congested.	Plays higher postpone performance, and therefore, a quicker dealing with of regulations installation in contrast with centralized FMCC architecture
63	2018	Efficient heuristic algorithm called MuSIC.	Function redundancy, system complexity, and inefficiency	73% of optimal, 30% better than simple strategies.
64	2016	Virtual machine migration approach.	Data traffic and the interference.	Limit the energy consumption at the facts center.
65	2016	Intra-PAN mobility management scheme.	No longer scale to cowl the destiny wishes of cell customers, lack of latency.	Reduces handoff delay, it minimizes the hyperlink disconnection time of MNS and packets loss during their handoffs.
66	2017	Unified offloading mechanism.	Battery life of mobile device is restricted, mobility performance is also limited.	Reduced amount of time.
67	2017	Distributed mobility management.	Cost is high, operating frequency is up to 2.66-ghz, several pcs or a high performance	Lessen the revenue- outstripping fees whilst stretching their community capacities with information offloading technology.
68	2017	Hierarchical resource management scheme for MCC.	Cost implications of this technique is high.	Energy can be drastically reduced.
69	2017	They proposed architecture to defend the cloudlet servers from viable DDOS attacks.	Limit the mission execution and records switch time to the consumer, large-scale datacenter networks which consist of more than one dispersed server centers.	Lessen the capital expenditure (CAPEX) and operational expenditure (OPEX), permit short failure recovery, and simplify the installation/modification of latest offerings on the SDN controller.
70	2018	Mobility-conscious best aid allocation architecture, specifically mobi-het.	Offset the potential data loss caused by node mobility.	Better performance in timeliness and reliability.

III. Problem Methodology and Proposed System Model

A. Problem Methodology

Aissioui et al. [71] have foreseen a versatile technique which is gotten from SDN/Open-Flow auxiliary plan and a control plane structure. They are commonly adjusted for versatile distributed computing frameworks and Follow Me Cloud (FMC)- related frameworks. The condition of 5G portable systems are predominantly utilizes to deal with the plane structure for MCC-related frameworks where versatile hubs and system administrations are demonstrating restriction of exercises and movements. Also, the disparity of unified strategy through lone SDN controller is encouraging to dispense the SDN/Open-Flow control plane on a two-level various leveled auxiliary structure which are containing first stage among a worldwide controller of G-FMCC and second stage among various neighborhood controller. The appraisal outcomes are gained by methods for examination. Also, this clarification is ensuring the upgraded control plane association, introduction support, and system asset protection. MCC is a focalized innovation included three foundation heterogeneous advancements, to be specific portable processing, distributed computing, and systems administration. The forthcoming heterogeneous 5G arrange underscores on an emotional increment in the transmission pace of MCC traffic. With more clients working at high rates, the sort of information shared over the system will be intricate and a dominant part of it will incorporate video traffic. Such mind boggling structure of traffic and substantial burden over the parts of the system are hard to control. Further, the portability of clients indicates this issue and makes it hard to oversee and work the system with no breakdown. In this manner, it is essential to control traffic just as deal with the versatility of clients to give effective correspondence, which can bolster video traffic at high conveyance rates. Also, the essential parameters influence the presentation of system as the tremendous assortment of cell phones with various Operating Systems (OSs), stages, and remote system measures. The up and coming 5G systems target giving rapid correspondences to clients independent of their development. With an expansion in the quantity of gadgets and the system achieving its pinnacle size, because of thick sending, it gets essential to oversee and control versatility for effective correspondence. Portability the executives requires various activities at a similar example, which incorporate ideal course choice, versatile stay support, client design distinguishing proof, and administration handoffs [72].

An energy efficient mobility management in mobile cloud computing (E2M2C2) system utilizes the optimal cluster selection, cloud selection and route selection to obtain energy efficient mobility management. The main involvement of proposed E2M2C2 system is summarized as follows:

- In E2M2C2 system, the Cluster Head Selection (CHS) Algorithm is used for reducing energy consumption in networking and data transmission phase.
- Various parameters used for cluster head selection process are: energy consumption, delay, handoff, overhead, delivery ratio.
- An Elective Repeat Multi-objective Optimization (ERMO^2) Algorithm is used to compute the best cloud among various in the network.
- Various parameters used for best cloud election process are: delay, packet loss rate, energy consumption, through put and fairness index.
- The Back Track Searching (BTS) Algorithm used to compute the congestion and select optimal routes between the serving terminals.
- Various parameters used for best cloud election process are: delay, packet loss rate, energy consumption, through put and fairness index.

B. System Model of Proposed E2M2MC2 System

The system model of proposed E2M2C2 framework is appears in Fig. 3, which utilizes the follow me cloud (FMC) idea, which permits the migration of administrations gave to clients contingent upon their developments. Administrations are along these lines consistently gave from server farm areas that are ideal for the present areas of the clients. This furnishes clients with improved QoS/QoE, simultaneously and it permits saving administrators' system assets by getting away system traffic to server farms through the closest focuses contrasted and clients' areas. Another bit of leeway of FMC innovation is that movement of administrations is consistent and straightforward to clients. MCC system utilize both the data storing and the data dealing with occur outside of the mobile phone. Concerning definition, portable applications move the preparing power and limit from the phone phones to the Cloud. It may be thought the union of the distributed computing and versatile condition [72].

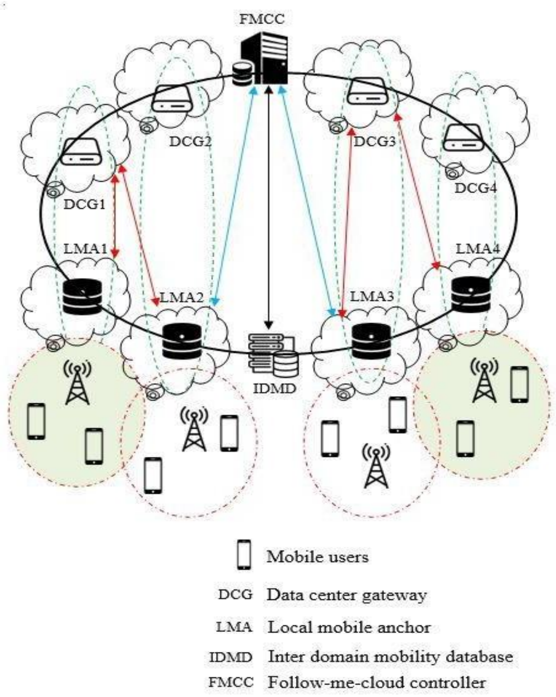


Fig. 3. System Model of Proposed E2M2MC2 System Block Diagram.

1) Cloud clusters group optimal algorithm: Clustering is a scarce and non-renewable is energy efficient design and the performance analysis requires proper model for measure proper energy consumption of network interfaces. The possible states of energy consumption in mobile nodes are transmit, receive, idle and sleep. The first two sate are when the mobile nodes is transmitting/receiving packets respectively, the idle state is when the mobile node is waiting for transmitting packets and the sleep state can neither transmit/receive packets. The link-metric cost (L_c) associated with the each packet at each mobile node is defined as the total incremental cost (C_t) proportional to the packet size (P_s) and a fixed cost (C_f) associated with the routing.

$$L_c = C_t \times P_s \times C_f \tag{1}$$

The average energy consumption (E_a) is written as follows:

$$E_a = a\,L_c\,P_a + b\;L_c\,P_s \tag{2}$$

where P_a is the power consumed by the mobile node with the cost of L_c and 'a' is the rate of occurrence, and P_s that is the power consumed by the mobile node with cost L_c and has the occurrence rate 'b'. Given the diversity of energy collection methods, and the wide range of application profiles it is not possible to create a generic model, however, the essential criterion is that the energy stored (E_{st}) in the node must be at least equal with the energy used for its operation in the time interval $T_2 - T_1$.

$$E_{st} = \int_{T_1}^{T_2} (P_c - P_{cs})dt \tag{3}$$

where P_c is the power consumed by the mobile node in the time interval $T_2 - T$ and P_{cs} is the power collected and stored power in the same timeline.

General building of mobile nodes arranged around a single power supply, which for the periods in which handset, when they are either ended by electronic switches, or set into rest state, it is ideally to be set to make a lower yield voltage through component vapor sorption strategy in light of the fact that the imperativeness viability of the microcontroller will increase in the midst of these seasons of rest states. The total energy consumed (Total $_{EC}$) by mobile node will be represented as follows:

$$Total_{EC} = \sum_{k=0}^{T} \frac{(E_a(t) + E_{st}(t))L_c}{E_{\eta_{DC-DC}}} \tag{4}$$

where $E_{\eta_{DC-DC}}$ is the energy consumption of DC-DC converter.

Once the distance between two neighborhoods exceed a certain extent, the transmission signal will not be received correctly by receiver i.e. link failure. The received signal strength at any node from its neighbor node affects. If a node receives a strong signal from a neighbor then the link between them is considered as stable, otherwise the link is considered as unstable. The difference of two signal strengths received at a node at two different times. The signal strength is stronger, it means that two nodes would be closer and the link between them would have longer lifetime. It is calculated in order to determine whether the node is within the transmission range or not. There are three main radio propagation models are free space, two-ray ground reflection and shadowing model. Simply, the received signal strength is defined as a ratio of the received power (P_r) to the reference power (P_{ref}).

$$R = 10\log\frac{P_r(x)}{P_{ref}} \tag{5}$$

The configured transmission power at the transmitting device (P_t) in total energy consumption is directly affects the receiving power at the receiving device (P_r).

$$P_r = P_t\,L_c\,\frac{\lambda^2}{4\pi\,d^2} \tag{6}$$

where, d is the distance between transmitter and receiver mobile node; λ is the control parameter taken from the mobility of mobile nodes.

2) Cloud selection using elective repeat multi-objective optimization (ERMO^2) algorithm: The below Fig. 4 flowchart describes the flow of ERMO^2 algorithm.

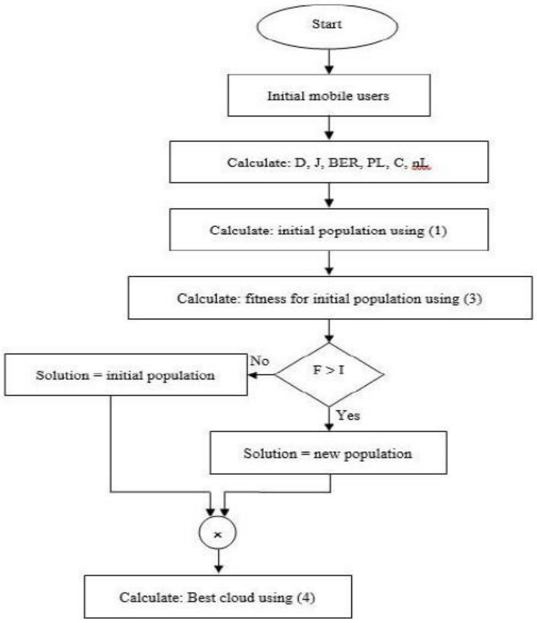

Start
Initial mobile users
Calculate: D, J, BER, PL, C, nL
Calculate: initial population using (1)
Calculate: fitness for initial population using (3)
No
F > I
Solution = initial population
Yes
Solution = new population
×
Calculate: Best cloud using (4)

Fig. 4. Working Flow of Proposed E2M2C2 System-Cloud Selection using ERMO^2 Algorithm.

IV. Result Analysis

The mobile cloud computing structures is dissected for the presentation perspective. Versatile cloud design builds asset accessibility by utilizing enormous number of close by cell phones in broad daylight places like shopping center, film, and air terminal assistance accessibility is expanding recognizably. It additionally upgrades security as a result of the dynamic parceling of the correspondence divert in the between cell phone and cloud server. In this segment, we present the assessment of our energy efficient mobility management in mobile cloud computing (E2M2C2) system and it compared with the existing distributed follow me cloud controller (DFMCC). The presentation of proposed E2M2C2 framework is investigated by the diverse testing situations: effect of versatile client thickness and their speed. The quantity of versatile clients is fluctuated from 30 to 110 in first test and the portable client speed is changed from 20 to 100 ms in the subsequent situation. For this testing, we utilize four DCG and LMA, one IDMD and one FMCC with high thickness portable clients. The two tests are actualized in Network Simulator (NS2) device with 1000×1000 m2 arrange size. The recreation parameters are condensed in Table VI.

TABLE. VI. Simulation Parameters

Parameters	Values
Number of Mobile nodes	30, 50, 70, 90, 110
Mobile user speed (ms)	20, 40, 60, 80, 100
Number of DCG	4
Number of IMA	4
Number of IDMD	1
Number of FMCC	1
Network size	1000×1000 m2
Traffic model	Constant bit rate
Simulation time (s)	100

1) *Effect of mobile nodes:* In this situation, we fluctuating the quantity of hub from 30 to 110 with the fixed speed as 60 ms and the presentation of proposed E2M2C2 is contrasted and the current DFMCC framework. Fig. 5 shows the bundle misfortune pace of proposed E2M2C2 and existing DFMCC framework. The plot obviously delineates the parcel misfortune pace of proposed E2M2C2 framework is exceptionally low regarding 41% less contrasted with existing DFMCC framework. Fig. 6 shows the vitality utilization of proposed E2M2C2 and existing DFMCC framework. The plot obviously delineates the vitality utilization of proposed E2M2C2 framework is low as far as 39% less contrasted with existing DFMCC framework. Fig. 7 shows the throughput of proposed E2M2C2 and existing DFMCC framework. The plot obviously delineates the throughput of proposed E2M2C2 framework is extremely high as far as 24% high contrasted with existing DFMCC framework. Fig. 8 shows the reasonableness list of proposed E2M2C2 and existing DFMCC framework. The plot obviously delineates the decency record of proposed E2M2C2 framework is high regarding 44% high contrasted with existing DFMCC framework. Fig. 9 shows the deferral of proposed E2M2C2 and existing DFMCC framework. The plot unmistakably portrays the deferral of proposed E2M2C2 framework is low as far as 20% high contrasted with existing DFMCC framework.

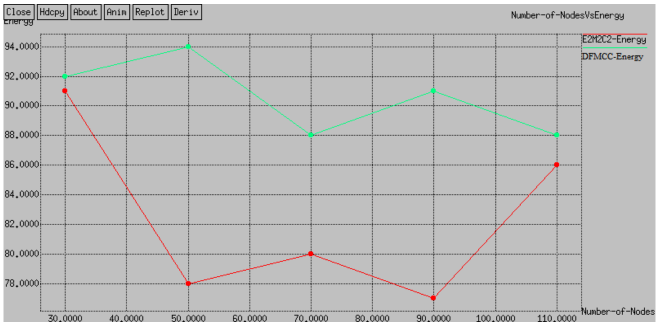


Fig. 5. Energy Consumption Comparison with Effect of Mobile Nodes.

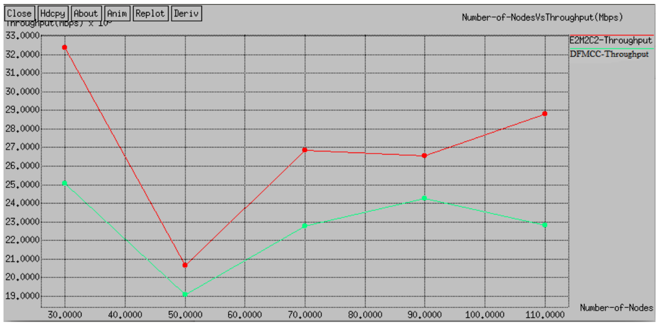


Fig. 6. Throughput Comparison with Effect of Mobile Nodes.

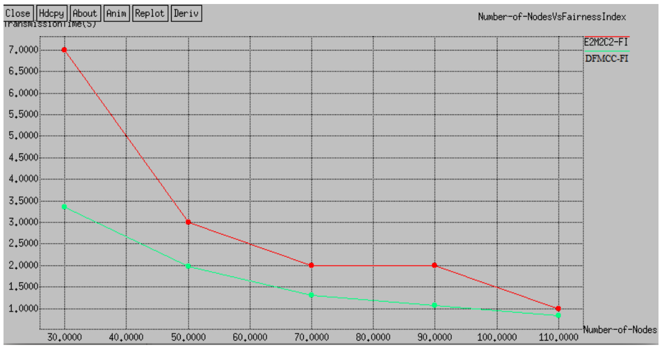


Fig. 7. Fairness Index Comparison with Effect of Mobile Nodes.

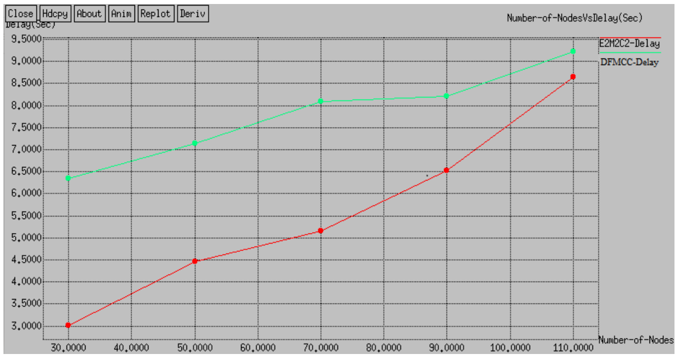


Fig. 8. Delay Comparison with Effect of Mobile Nodes.

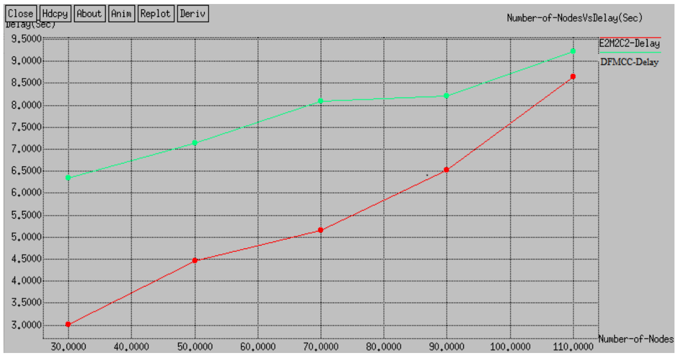


Fig. 9. Delay Comparison with effect of Mobile Nodes.

V. Conclusion

In this paper, we have embraced an efficient writing survey of versatility the executives in MCC, so as to comprehend the pattern of research interests so far in MCC, regarding the least and most inquired about issues. We had the option to feature a portion of the difficulties in MCC, for example, protection, security and trust, adaptation to non-critical failure, portability the executives, arrange clog, heterogeneity and association conventions, asset limitation and stage heterogeneity, setting mindfulness, introduction and convenience issues, battery life and vitality mindfulness, and cloud API Security Management. From these surveys, we concluded that the MCC framework is for the most part influenced by the portability. Consequently, proficient systems are required for additional upgrade in group of people yet to come systems.

References

[1] Linthicum, "Connecting Fog and Cloud Computing," in IEEE Cloud Computing, 2017, vol. 4, no. 2, pp. 18- 20.

[2] Linthicum, D.S., "The Technical Case for Mixing Cloud Computing and Manufacturing," in IEEE Cloud Computing, 2016, 3(4), pp.12-15.

[3] Linthicum, D.S., "Cloud Computing Changes Data Integration Forever: What's Needed Right Now," in IEEE Cloud Computing, 2017, 4(3), pp.50-53.

[4] Zhou, B., Dastjerdi, A.V., Calheiros, R.N., Srirama, S.N. and Buyya, R., " mCloud: A context-aware offloading framework for heterogeneous mobile cloud," in IEEE Transactions on Services Computing, 2017, 10(5), pp.797- 810.

[5] Fowley, F., Pahl, C., Jamshidi, P., Fang, D. and Liu, X., "A classification and comparison framework for cloud service brokerage architectures," in IEEE Transactions on Cloud Computing, 2018, 6(2), pp.358-371.

[6] Shah, S.C., "Recent advances in mobile grid and cloud computing," in Intelligent Automation & Soft Computing, 2017, pp.1-13.

[7] Fu, J., Jones, M., Liu, T., Hao, W., Yan, Y., Qian, G. and Jan, Y.K., "A novel mobile-cloud system for capturing and analyzing wheelchair maneuvering data: a pilot study," in Assistive Technology, 2016, 28(2), pp.105-114.

[8] Xu, B., Xu, L., Cai, H., Jiang, L., Luo, Y. and Gu, Y.,. "The design of an m-Health monitoring system based on a cloud computing platform," in Enterprise Information Systems, 2017, 11(1), pp.17-36.

[9] Tseng, F.H., Cho, H.H., Chang, K.D., Li, J.C. and Shih, T.K.,. "Application-oriented offloading in heterogeneous networks for mobile cloud computing," in Enterprise Information Systems, 2018, 12(4), pp.398-413.

[10] Morales-Sandoval, M., Vega-Castillo, A.K. and Diaz-Perez, A., "A secure scheme for storage, retrieval, and sharing of digital documents in cloud computing using attribute-based encryption on mobile devices," in Information Security Journal: A Global Perspective, 2014, 23(1-2), pp.22-31.

[11] Singh, R., "Genetic-variable neighborhood search with thread replication for mobile cloud computing," in International Journal of Parallel, Emergent and Distributed Systems, 2017, 32(5), pp.486-501.

[12] Psannis, K.E., Xinogalos, S. and Sifaleras, A.,. "Convergence of Internet of things and mobile cloud computing. Systems Science & Control Engineering," in Open Access Journal, 2014, 2(1), pp.476-483.

[13] Xu, G., Yu, W., Chen, Z., Zhang, H., Moulema, P., Fu, X. and Lu, C.,. "A cloud computing based system for cyber security management," in International Journal of Parallel, Emergent and Distributed Systems, 2015, 30(1), pp.29-45.

[14] Mamei, M., Roli, A. and Zambonelli, F.,. "Emergence and control of macro-spatial structures in perturbed cellular automata, and implications for pervasive computing systems," in IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2005, 35(3), pp.337-348.

[15] Samimi, F.A., McKinley, P.K., Sadjadi, S.M., Tang, C., Shapiro, J.K. and Zhou, Z.,. "Service clouds: distributed infrastructure for adaptive communication services," in IEEE Transactions on Network and Service Management, 2007, 4(2), pp.84-95.

[16] Lin, Y., Shao, L., Zhu, Z., Wang, Q. and Sabhikhi, R.K., "Wireless network cloud: Architecture and system requirements," in IBM Journal of Research and Development, 2016, 54(1), pp.4-1.

[17] Kumar, K. and Lu, Y.H., "Cloud computing for mobile users: Can offloading computation save energy?" in Computer, 2010, 43(4), pp.51-56.

[18] Li, Z., Chen, C. and Wang, K., "Cloud computing for agent-based urban transportation systems," in IEEE Intelligent Systems, 2010, 26(1), pp.73-79.

[19] Deboosere, L., Vankeirsbilck, B., Simoens, P., De Turck, F., Dhoedt, B. and Demeester, P., "Cloud-based desktop services for thin clients," in IEEE Internet Computing, 2012, 16(6), pp.60-67.

[20] Ahmed, A., Boulahia, L.M. and Gaiti, D., "Enabling vertical handover decisions in heterogeneous wireless networks: A state-of-the-art and a classification," in IEEE Communications Surveys & Tutorials, 2014, 16(2), pp.776-811.

[21] Tian, D., Wei, J., Zhou, J., Sheng, Z., Chen, M., Ni, Q. and Leung, V.C., "From cellular decision making to adaptive handoff in heterogeneous wireless networks," in IEEE Wireless Communications Letters, 2018, 7(1), pp.2- 5.

[22] Deb, S. and Monogioudis, P., "Learning-based uplink interference management in 4G LTE cellular systems," in IEEE/ACM Transactions on Networking (TON), 2015, 23(2), pp.398-411.

[23] Wang, F., Wang, Z., Qian, C., Dai, L. and Yang, Z., "Efficient vertical handover scheme for heterogeneous VLC-RF systems," in Journal of Optical Communications and Networking, 2015, 7(12), pp.1172-1180.

[24] Nguyen-Vuong, Q.T., Agoulmine, N., Cherkaoui, E.H. and Toni, L., "Multicriteria optimization of access selection to improve the quality of experience in heterogeneous wireless access networks," in IEEE Transactions on Vehicular Technology, 2013, 62(4), pp.1785-1800.

[25] Mehmeti, F. and Spyropoulos, T., "Performance analysis of mobile data offloading in heterogeneous networks," in IEEE Transactions on Mobile Computing, 2017, 16(2), pp.482-497.

[26] La Corte, A., A. Puliafito, and O. Tomarchio, "QoS management in programmable networks through mobile agents," in Microprocessors and Microsystems, 2001, 25(2): p. 111-120.

[27] Manner, J., et al., "Evaluation of mobility and quality of service interaction," in Computer Networks, 2002, 38(2): p. 137-163.

[28] Aoul, Y.H., et al. "M3G: A mobile multicast multimedia gateway for seamless IPv4/IPv6 transition," in IFIP/IEEE International Conference on Management of Multimedia Networks and Services, 2003, Springer.

[29] Mun, Y. and Y. Kim, "A location management scheme to provide IP mobility over wireless ATM," in Future Generation Computer Systems, 2004, 20(2): p. 205-219.

[30] Kallath, D., "Trust in trusted computing – the end of security as we know it" in Computer Fraud & Security, 2005, (12): p. 4-7.

[31] De la Oliva, Antonio, Carlos Jesús Bernardos, and María Calderón. "Practical evaluation of a network mobility solution." In EUNICE 2006:

Networks and Applications Towards a Ubiquitously Connected World. Springer, Boston, MA, 2006, 133-144.

[32] Samimi, Farshad A., et al. "Service clouds: distributed infrastructure for adaptive communication services," in IEEE Transactions on Network and Service Management 4.2, 2007, 84-95.

[33] Harada, Hiroshi, et al. "A software defined cognitive radio system: cognitive wireless cloud," in Global Telecommunications Conference, 2007, GLOBECOM'07, IEEE.

[34] Munasinghe, Kumudu S., and Abbas Jamalipour. "A 3GPP-IMS based approach for converging next generation mobile data networks Communications", in ICC'07. IEEE International Conference on. IEEE, 2007.

[35] Varga, E., et al. "Novel information dissemination solutions in biologically inspired networks Telecommunications," in ConTel 2007. 9th International Conference on. IEEE, 2007.

[36] Xiao, Yunpeng, Yang Tao, and Qian Li. "A new wireless web access mode based on cloud computing," in Computational Intelligence and Industrial Application, 2008. PACIIA'08. Pacific-Asia Workshop on. Vol. 1. IEEE.

[37] Ghassemian, Mona, and Hamid Aghvami. "An investigation of the impact of mobility on the protocol performance in wireless sensor networks," in Communications, 2008 24th Biennial Symposium on. IEEE, 2008.

[38] Ahmad, Syed Zubair, Muhammad Abdul Qadir, and Mohammad Saeed Akbar. "A distributed resource management scheme for load-balanced QoS provisioning in heterogeneous mobile wireless networks," in Proceedings of the 4th ACM symposium on QoS and security for wireless and mobile networks. ACM, 2008.

[39] Boldrini, Chiara, Marco Conti, and Andrea Passarella. "Exploiting users' social relations to forward data in opportunistic networks: The HiBOp solution," in Pervasive and Mobile Computing , 2008, 4.5 (2008): 633-657.

[40] Toor, Yasser, et al. "Vehicle ad hoc networks: Applications and related technical issues," in IEEE communications surveys & tutorials 10.3, 2008, 74-88.

[41] Shaw, Gareth, and Allan Williams. "Knowledge transfer and management in tourism organisations: An emerging research agenda," in Tourism Management 30.3, 2009, 325-335.

[42] [Kim, Jonghyun, Vinay Sridhara, and Stephan Bohacek. "Realistic mobility simulation of urban mesh networks," in Ad Hoc Networks 7.2, 2009, 411-430.

[43] [Elston, Jack, et al. "Net-centric communication and control for a heterogeneous unmanned aircraft system," in Journal of intelligent and Robotic Systems 56.1-2, 2009, 199-232.

[44] Klein, Andreas, et al. "Access schemes for mobile cloud computing," in Mobile Data Management (MDM), 2010 Eleventh International Conference on. IEEE, 2010.

[45] Liu, Leslie, Randy Moulic, and Dennis Shea. "Cloud service portal for mobile device management," in e- Business Engineering (ICEBE), 2010 IEEE 7th International Conference on. IEEE, 2010.

[46] Satyanarayanan, Mahadev. "Mobile computing: the next decade," in Proceedings of the 1st ACM workshop on mobile cloud computing & services: social networks and beyond. ACM, 2010.

[47] Huerta-Canepa, Gonzalo, and Dongman Lee. "A virtual cloud computing provider for mobile devices," in Proceedings of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond. ACM, 2010.

[48] Munasinghe, Kumudu S., and Abbas Jamalipour. "An analytical evaluation of mobility management in integrated WLAN-UMTS networks," in Computers & Electrical Engineering 36.4, 2010, 735-751.

[49] Antoniou, Josephine, et al. "Supporting context-aware multiparty sessions in heterogeneous mobile networks," in Mobile Networks and Applications 15.6, 2010, 831-844.

[50] Armbrust, Michael, et al. "A view of cloud computing," in Communications of the ACM 53.4, 2010, 50-58.

[51] Li, ZhenJiang, Cheng Chen, and Kai Wang. "Cloud computing for agent-based urban transportation systems," in IEEE Intelligent Systems 26.1, 2011, 73-79.

[52] Sardis, Fragkiskos, et al. "On the investigation of cloud-based mobile media environments with service- populating and QoS-aware mechanisms," in IEEE transactions on multimedia 15.4, 2013, 769-777.

[53] Amoroso, Edward G. "From the enterprise perimeter to a mobility-enabled secure cloud," in IEEE Security & Privacy 11.1 (2013): 23-31, 2013.

[54] Panta, Rajesh K., et al. "Phoenix: Storage using an autonomous mobile infrastructure," in IEEE Transactions on Parallel and Distributed Systems 24.9, 2013, 1863-1873.

[55] Gkatzikis, Lazaros, and Iordanis Koutsopoulos. "Migrate or not? exploiting dynamic task migration in mobile cloud computing systems," in IEEE Wireless Commun. 20.3, 2013, 1-0.

[56] Taleb, Tarik, Konstantinos Samdanis, and Adlen Ksentini. "Supporting highly mobile users in cost-effective decentralized mobile operator networks," in IEEE Transactions on Vehicular Technology 63.7, 2014, 3381-3396.

[57] Zhang, Daqiang, et al. "Mobility prediction in telecom cloud using mobile calls," in IEEE Wireless Communications 21.1, 2014, 26-32.

[58] Qi, Qi, Jianxin Liao, and Yufei Cao. "Cloud service-aware location update in mobile cloud computing," in IET Communications 8.8, 2014, 1417-1424.

[59] Chen, Shanzhi, et al. "Mobility-driven networks (MDN): from evolution to visions of mobility management," in IEEE Network 28.4,2014, 66-73.

[60] Yue, Hao, et al. "DataClouds: Enabling community-based data-centric services over the Internet of Things," in IEEE Internet of Things Journal 1.5, 2014, 472-482.

[61] Li, Wenzhong, et al. "Mechanisms and challenges on mobility-augmented service provisioning for mobile cloud computing," in IEEE Communications Magazine 53.3, 2015, 89-97.

[62] Aissioui, Abdelkader, et al. "Toward Elastic Distributed SDN/NFV Controller for 5G Mobile Cloud Management Systems," in IEEE Access 3.0, 2015, 2055-2064.

[63] Rahimi, M. Reza, et al. "On optimal and fair service allocation in mobile cloud computing," in IEEE Transactions on Cloud Computing 2015.

[64] Kumar, Neeraj, Sherali Zeadally, and Joel JPC Rodrigues. "Vehicular delay-tolerant networks for smart grid data management using mobile edge computing," in IEEE Communications Magazine 54.10, 2016, 60-66.

[65] Ha, Minkeun, Seong Hoon Kim, and Daeyoung Kim. "Intra-MARIO: A Fast Mobility Management Protocol for 6LoWPAN," in IEEE Transactions on Mobile Computing 16.1, 2017, 172-184.

[66] Ha, Minkeun, Seong Hoon Kim, and Daeyoung Kim. "Intra-MARIO: A Fast Mobility Management Protocol for 6LoWPAN," in IEEE Transactions on Mobile Computing 16.1, 2017, 172-184.

[67] Jeon, Seil, et al. "Distributed Mobility Management for the Future Mobile Networks: A Comprehensive Analysis of Key Design Options," in IEEE Access 5, 2017, 11423-11436.

[68] Ahmad, Awais, et al. "Energy efficient hierarchical resource management for mobile cloud computing," in IEEE Transactions on Sustainable Computing 2.2, 2017, 100-112.

[69] Chaudhary, Rajat, Neeraj Kumar, and Sherali Zeadally. "Network service chaining in fog and cloud computing for the 5G environment: Data management and security challenges," in IEEE Communications Magazine 55.11, 2017, 114-122.

[70] Enayet, Asma, et al. "A mobility-aware optimal resource allocation architecture for big data task execution on mobile cloud in smart cities," in IEEE Communications Magazine 56.2, 2018, 110-117.

[71] A. Aissioui, A. Ksentini, A. Gueroui and T. Taleb, "Toward Elastic Distributed SDN/NFV Controller for 5G Mobile Cloud Management Systems", IEEE Access, vol. 3, pp. 2055-2064, 2015.

[72] L. Pallavi, A. Jagan, B. Thirumala Rao, "ERMO2 algorithm: an energy efficient mobility management in mobile cloud computing system for 5G heterogeneous networks", IJECE, vol. 9, 2019.

State-of-the-Art Reformation of Web Programming Course Curriculum in Digital Bangladesh

Susmita Kar[1], Md. Masudul Islam[2], Mijanur Rahaman[3]
Department of CSE
Bangladesh University of Business and Technology (BUBT)
Dhaka, Bangladesh

***Abstract*—For last 15 years universities around the world are continuously developing effective curricula for Web Engineering in order to create good opportunities for graduates to cope up with IT-Software industries. From this study we will show the gap between the skill requirements of IT-Software industries and universities' web course curricula. Also, we will provide a balanced and structured web course curriculum for any universities. Nowadays, there is a rapid development in web-based applications everywhere but most of our students are late bloomer in programming. So, to ease their difficulties in web sector we need a balanced web curriculum and effective teaching method. By this curriculum one can achieve an overall idea and a minimum view of web engineering which can be beneficial for them in further Web development. Students get a little knowledge in their university on Web Engineering because of the vastness of the contents and the small duration of semester. Our two-semester web course curricula will help them to overcome this problem. Two-semester web course curricula have a huge impact on achieving the minimum required skill in web development field in IT-Software industries. It will help to obtain most of the area of web related content also it will increase problem solving skill and versatile knowledge of web engineering in undergraduate life.**

Keywords—Web engineering; web development; outcome based learning; CDIO; web course curriculum; web ecosystem; digital Bangladesh

I. Introduction

The world is in hype with internet and its broad technology. Almost every applications and digital concern is now web-based. Even, with the rapid development of internet most of the software system is now converting their services into web applications. If you want to develop any software, turn it into web system for accessibility. The capability of developing web applications has become a must-have professional skill for IT areas, exclusively for those who are graduates.

Since Digital Bangladesh program was launched in 2009 the government has been proactively chasing the digital wings in a full throttle. Aiming at transforming Bangladesh into digital economy by 2021 and knowledge-based economy in 2041 billion dollars projects are ongoing. Within 2023 almost 45000+ government websites will be published for services. Also, we are introducing our IT expertise in a billion-dollar market place of IT industries [1].

However, this rapid changes and development requires a lot of engineering and merits. This is why many of our universities, institutes and training centers are creating awareness of importance of web application development. Offering web engineering courses is now a trend and must-have thing in computer science. The ACM and IEEE Computer Society have added web programming courses as electives to the CS curriculum in the new CS2013 [2]. Web programming has become a dominant programming model. Still in our country most of the curriculum seems like backdated and weak. In the beginning of 2006, the traditional computer science curriculum did not include web engineering courses [3]. Gradually, many web courses are being introduced in different semesters in various institutes' curriculum according to that generation's advancements. These courses are all about some web programming techniques, programming languages, procedures and web database driven applications. It appeared to be clear that, to develop web applications multiple levels of skill is required. But recent survey shows that, there are some standards and professional view of industries which are absent in our present universities' curriculum. These curricula and course contents should be made according to the requirements of updated web ecosystem. So, many changes are introduced in recent days to develop professionalism in web engineering but we need proper guidance from beginning of undergraduate life. The growing importance of web application development on computing education is a proof that we need a concrete web curriculum in university studies.

Therefore, this paper will have three-fold: 1) Summarize the survey result of recent IT-Software Job trends. 2) Academic curriculum status of web engineering in different universities in Bangladesh. 3) A proposed outcome-based Web course design with effective sociocultural teaching method.

II. Related Works in Developing and Teaching Web Programming Course

In the beginning, computer science study didn't give enough priority in web engineering. Most of the world wide universities had no web courses except giving some basic knowledge of internet. Since 1990, Tampere, a web programming course, has been taught in University of Technology and that was updated in 2010 [4]. The University of Texas Pan American decided to offer web programming course from 2005 for undergraduates [3]. A simple web programming course for students was developed by Stepp, Miller, and Kirst which required no prerequisites [5]. Robert E. Noonan implemented an advanced web programming course that highlighted server-side programming, database interaction, and security [6]. Xusheng Wang updated a web programming course which is based on server-side techniques with PHP and MySQL [7] and in 2014 he proposed second web programming course which contains Web 2.0 technologies and CMS [8]. M. J. Lantis in his paper presented a web editor as a development platform to teach HTML and client-side programming [9]. In 2011, Connolly proposed three-semester web course sequences which contained HTML, CSS, JavaScript, server-side programming, server-side frameworks, user experience, security, deployment, hosting, and web services topics [10]. But Robert revised the idea and made a single semester course model [11]. Francesco Maiorana from University of Catania designs a curriculum suited both for graduates and for a third-level high school web programming course [12].

Based on different case studies and teaching experiences on web programming at undergraduate studies we can list the following issues where Liu in 2011 presented challenges and tools used [13], Laverty in 2011 point-outs the difficulties of an efficient delivery of a dynamic web development, database-driven platform [14] Baatard in 2007, offered a course using the PHP language with security problems [15]. Moreover, Noonan in 2007, Wang in 2006 and Olan in 2009 announced a course fixated on server-side programming and database interaction [6][7][16]. Gousie in 2006 shows on an interdisciplinary method to teach web programming, graphics and design in a course [17]. Stepp in 2009, presented a web programming course suggested to instructors to introduce PHP from the beginning [5] and Adams in 2007 developed a Web project-oriented course [18]. Recently Chao in 2013, reviewed the usage of framework but which is not suggested in preliminary web programming courses [19]. These are the summary view of ongoing activities and suggestions of updating web curricula in various universities to achieve standard goal towards web development in undergraduate level.

III. Motivation

Web Engineering or Web development is a huge task. Various types of knowledge, tools and techniques are needed to complete this engineering. Several facts we need to consider for web programming are: analysis, design, UX/UI, backend system, framework, database, security and maintenance. But the main issue that we should consider is to develop practical web application for industries that could benefit our country because day by day our country is adapting digital technology and online services. That is why we will emphasis on academic view of web development and their pros and cons.

IV. State-of-the-Art in IT-Software Job Field in Bangladesh

In order to keep pace with recent trends, we need to find out the updated requirements of IT-Software job fields in our country. We have surveyed last two years' job requirements in IT industries.

A. *Inspection on Job Skill Requirements*

The survey contains 2018 and 2020's data from country's leading job site bdjobs.com. Basically, we searched for IT/Engineering category jobs and among them we took 150+ valid job circulars around the country.

We have filtered the specific web development jobs from hundreds of jobs. We found a total of 163 valid IT Jobs in September 2018 and 180 valid IT jobs in January 2020 posted in website. From there we have collected requirements and additional skill requirements concerned with web engineering mostly. The following skills we fetched there: HTML, CSS, UX/UI, JavaScript, JSON, jQuery, Angular, VueJs, Various Scripting Language, PHP, API, PHP Framework, Laravel, ASP.Net, Web Engineering Concepts, Digital Marketing, Oracle, MySQL/MSSQL, System Analysis, Problem Solving, Project Management and Content Management/ Documentation. Almost in every job these skills are introduced as must-have issues for any candidate. In Fig. 1, the sorted data chart shows that from the end of 2018 to beginning of 2020 the IT related jobs have increased. Also, in each job the following skills are essential according to the view of employer. From the chart we can easily figure it out that, client-side language is leading the web-software industries then the server-side language is taking the role parallelly. The necessity of HTML, CSS, JS, PHP, MySQL, Framework, Web Database Programming, etc. proves that these contents should have serious impact on our academic curricula.

B. *Expert's Viewpoints*

To develop one's web development skills and keep growing with time so many expert communities provide voluntary advices and tutorials in open web. The most renowned community that created a tentative solid developer roadmap for beginner is roadmap.sh [20]. They have shown an infographic view to find a set of charts demonstrating the paths that anyone can take and the technologies that one would want to adopt in order to become a frontend, backend or a devops [20]. Summarized view of the roadmap-2020 is shown below in Table I.

There is more of it but in short, any beginner web developer can have a tentative idea of how he/she should move towards the world of web engineering. This is a vast journey and in this competitive market there is too short time for preparing oneself. That is why we want to make a well-developed syllabus for academic students so that they can have a strong base on web programming field after their graduation.

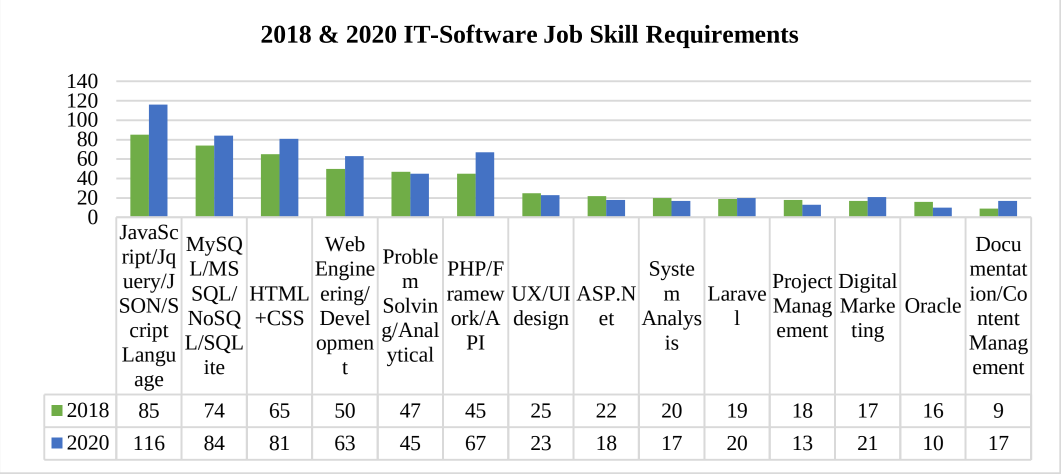


Fig. 1. 2018 & 2020 IT Industries Skill Requirements.

TABLE. I. ROADMAP 2020 FOR WEB DEVELOPER

Developer Type	Skills	Remarks
Front-End	Web Ecosystem, Internet and HTTP, Browser	*Recommended*
	HTML, Standard Practices	*Recommended*
	CSS, CSS3, Responsive Design	*Recommended*
	JavaScript, DOM, API, Event Handling, Client-side actions	*Recommended*
	Web Security, HTTPS, SSL	*Learn Anytime*
	Modern CSS, Web Components, Templates, CSS Framework, UI Design, Tools	*Better to Learn*
	Web Design Model, Server-side Rendering, Angular, Vue, React JS, GraphQL.	*Better to Learn*
	Version Control, Git	*Better to Learn*
	Mobile Application, Desktop Application, Web assembly	*Learn Anytime*
Back-End	Web Ecosystem, Internet and HTTP, Browser	*Recommended*
	Basic Front-End Knowledge, HTML, CSS, JS	*Recommended*
	Basic Networking Concepts and OS Knowledge	*Better to Learn*
	Java, PHP, C#, ASP.Net, Python	*Recommended*
	Version Control, Git	*Better to Learn*
	Relational Database, Query Language, SQL Engine, NoSQL, Normalization, Indexing	*Better to Learn*
	APIs	*Better to Learn*
	Web Security, Caching, Encryption, SSL	*Recommended*
	Unit, Functional and Integration Testing	*Better to Learn*
	Web and Database Server, Apache, IIS	*Recommended*
	Web Sockets	*Learn Anytime*

V. PRESENT CONDITION OF ACADEMIC CURRICULUM OF UNIVERSITIES OF BANGLADESH

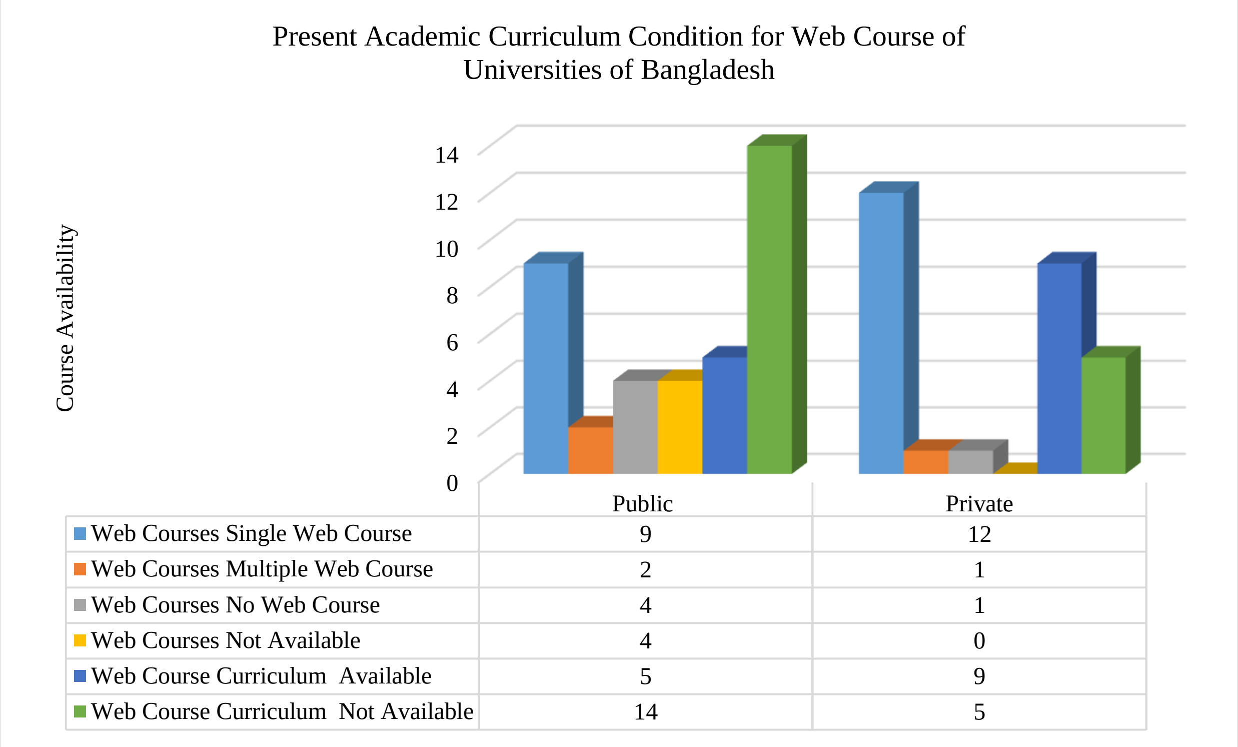
In 2020 we inspected more than 50 public and private universities which have engineering or IT related program. As per their current course curricula given in their official websites, we have collected 19 public and 14 private universities' course curricula of Computer Engineering or Information & Telecommunication Program.

A. Academic Inspection of Various Universities

Among these 33 universities here is the chart view of present condition according to the online survey shown in Fig. 2.

B. Limitations of Course Contents

According to our survey, we have some major findings on course syllabus of web programming. If we compare the available 14 curricula from both private and public universities the private universities' syllabus are richer and more updated. Public universities still contain generic and backdated contents in their course structure. Most of the web courses are introduced in 4th year and the basic prerequisite for the web courses are mostly programming and database. We have attached our findings in appendix [see Table IV in Appendix below].



	Public	Private
🟦 Web Courses Single Web Course	9	12
🟧 Web Courses Multiple Web Course	2	1
⬜ Web Courses No Web Course	4	1
🟨 Web Courses Not Available	4	0
🟦 Web Course Curriculum Available	5	9
🟩 Web Course Curriculum Not Available	14	5

Fig. 2. Present Academic Curriculum Condition for Web Course of Universities of Bangladesh.

VI. Issues and Challenges Towards the Web Ecosystem

Because of some limitations of old-style syllabus in computer engineering and IT-Software program, many undergraduate apprentices do not have a prospect to thoroughly learn web programming. Also, because of the high demands of job industries, many students are now leaning web programming. That is why improvised curriculum and learning method are needed to get something effective.

Nowadays, having simple static website is not enough for any company, institute or person. The success lies under the digital contents and the usage of modern tools and technologies within the website to make it riches in order to make anyone’s business successful. Hence, modern web ecosystem contains the followings: Content Management, Google analytics (or another analytics program), Clean, modern design using cascading style sheets for page layout, Secure login area (for updating content and for developing an intranet/extranet), Social media tie-ins to Facebook, linked in and twitter, etc. [21] As Vanessa Fox said, “Your online strategy is your business strategy”. Let us discuss some important web related issues that will clear our understanding that how it relates to our proposed web course syllabus. Website usability, a user investigation in 2004 (“Web usability - the main rules”, 2004) shows that about 40% of users never come back to a website after their first unsuccessful attempt [7]. One of the worst web designer’s mistakes is to create a website without a previous analysis of user needs [7]. Accessibility, the World Wide Web Consortium (W3C) defines a website as accessible if it allows access to people with disability (W3C, 2005). Privacy, the privacy of many web visitors is jeopardized due to data collecting on pages especially when the GET method of form submission is used, as it is known for its privacy vulnerabilities, and visitor tracking through cookies and web beacons [7]. Security, Hackers commonly use website vulnerabilities, and information related to them can be found on the Internet. Therefore, it is important to be informed about the vulnerabilities of company web application [7]. These are some fundamental issues that should be taken seriously during web development.

On the other hand, during the teaching web programming it faces many challenges. Web application is a complex, multi-faceted execution model, so focusing a single technology and performing a depth-oriented approach does not work for web programming [22]. For example, there are many design tools & technologies for instructor to teach front-end and back-end part such as HTML5, CSS3, Angular JS, Ruby on Rails, Bootstrap, Zend, Laravel, PHP, Python, API etc. It is tough for instructors to update their course contents always according to the rapid changes of web technologies and its features. The third challenge is inadequate integration among current web technologies and inconsistent implementation of standards [23]. Inconsistency between web browsers adds to the complexity of development, and these difficulties are reflected in the curriculum of web programming [24]. After reviewing the issues of web ecosystem and challenges of teaching web programming, we decided to focus on the objective to redesign the web course curriculum.

VII. Proposed Pedagogical Curriculum

According to ACM Curriculum Guidelines for Undergraduate Degree Programs in Computer Science 2013, there should be an elective web programming course which will cover basic web programming languages, some platform rule regulations, cloud technologies, following the web standards. After completion of the course one should be able to design and develop simple dynamic website with knowing proper constraints. They can also differ between software programming and web programming and identify how web standards impact on software development [25]. With all these things in mind we have designed two semester course.

Previously as we have seen recent job skill requirements in IT-Software industries it is clear that our Web programming syllabus must have two parts in a row: 1) Front-end part with templating; and 2) Back-end part as well as framework driven system.

So, we proposed two different courses in our academic curricula for web programming: 1. Web Programming I (Front-end) 2. Web Programming II (Back-end). We have shown the practical part of the course contents. Before introducing our course model, we need to understand what is Outcome-Based (OB) curriculum and then we will provide a well-developed web programming course syllabus which are mapped with outcome-based system because most of the university in our country are grabbing the Outcome-Based Education (OBE) system gradually to get international recognition.

A. Outcome based Web Curriculum

There are different aspects of Outcome Based (OB) curricula. The four-basic principle [26] of OBE are: Curriculum should have clear focus, designing curriculum with clear definition, Curriculum should arise high expectation of achievement, expand the opportunities for different learners. Based on these principles we propose two semester web programming courses. Each semester may be 4 months (max. 13 weeks) or 6 months (max. 19 weeks) in length but the core contents will be the same. Where course contents for 4 months' semester will be short and minimal on other hand content for 6 months' semester will be broad and elaborated.

Another important fact is due to the limitation of time duration, vastness and complexity of web contents. It's not necessary to cover all the topics in these two courses. Therefore, we design a syllabus that will enough to give an abstract idea of web application based on web ecosystem. We will follow the division technique shown in Fig. 3.

At a glance, in Table II, Web Programming-I curriculum (Front-End) part there will be basic idea and concepts for web 2.0 technologies, update design language HTML5, CSS3 and the most widely used and required scripting language JavaScript/jQuery. These contents should be taught and assign by problem-based strategy. After learning this tools student will be able to create single page static web design such as ID Card, single page resume and registration/login page with client-side validation in JS or online order form. They will investigate the problem and design proper solution through real life experiences. There is also a little bit of introduction of design framework such as Bootstrap, Material Design and responsive design in order to conceive the device-oriented design. Finally, the student will be given a small group project which contains usage of all these tools & technologies usages and show the demonstrations to their instructor. These curricula will cover a large area of their state-of-the-art of front-end development learning for students.

Moreover, In Table III, Web Programming-II (Back-End) Curriculum part there will be basic concepts of web server, installation, troubleshooting and security. Student will learn most widely used back-end language PHP and query language MySQL for a short period. As we know, CRUD (Create, Retrieve, Update & Delete) process is core of any web development, so we have added problem-based CRUD programming along with file upload and dynamic CRUD system in our syllabus. Model View Controller (MVC) model and latest prominent framework Laravel basic is also introduced in our syllabus. Finally, the students will make a back-end structure for their previously designed web project with simple admin panel where all of these back-end tools and technologies usage are met.

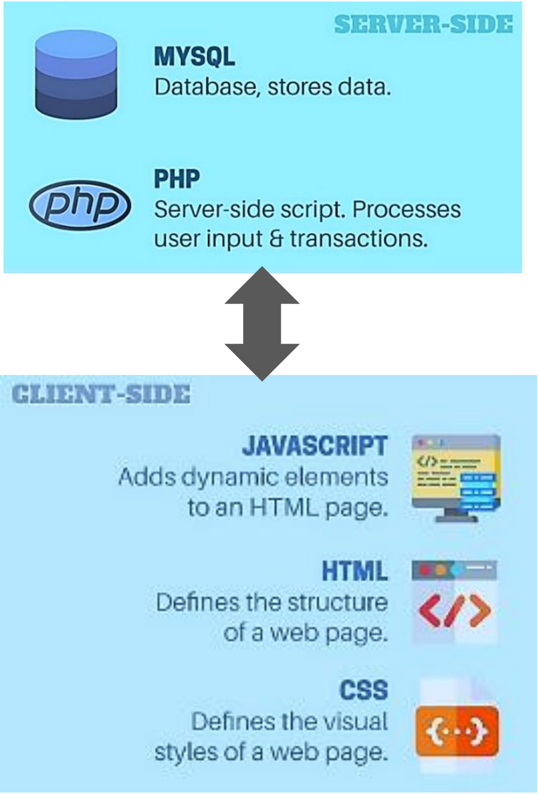

SERVER-SIDE
MYSQL
Database, stores data.
php
PHP
Server-side script. Processes user input & transactions.
CLIENT-SIDE
JAVASCRIPT
Adds dynamic elements to an HTML page.
HTML
Defines the structure of a web page.
CSS
Defines the visual styles of a web page.

Fig. 3. Web Application Division Technique.

B. Web Programming-I (Front-End)

TABLE. II. FIRST WEB COURSE SYLLABUS FOR FRONT-END PART

Week	Topics	Strategy & Learning Experience
1st	Web 2.0 fundamentals, HTTP, Web Standards, HTML, CSS, JS, DOM, Media Files, Browsers, IDE, Tools & Environmental setup.	Discussion, Overview & Concepts
2nd	HTML 5 elements, Attributes, styles, link, list, image, table, block, class, id, iframe, heading, comments.	Discussion, Assignment, Problem Based Learning
3rd	HTML 5 Basic Form design, Input, Radio button, Select Box, Checkbox, Files, Canvas, Media.	Discussion, Assignment, Problem Based Learning
4th	CSS 3 syntax, usage, Box Model, various selector, options, styling text, layout, typography.	Discussion, Assignment, Problem Based Learning
5th	CSS 3 Animations, effects, border, media types, responsive design	Discussion, Assignment, Problem Based Learning
6th	JavaScript basic syntax, Object, DOM, alert & dialogue box, various functions, events, Regular Expression, validations.	Discussion, Assignment, Problem Based Learning
7th	jQuery Basic, Object, Selector, Events, methods, API usage.	Discussion, Assignment, Problem Based Learning
8th	jQuery form validation, animations, Get, Set, Add, Remove, traversing.	Discussion, Assignment, Problem Based Learning
9th	Make Static Web Page using HTML5, CSS3(Personal single page CV, ID Card, Routine design),	Project & Inquiry Based Learning, Demonstration
10th	Make simple Form Design using HTML5, CSS3, JS validation (Online Registration Form, Online Login form, Pre-Book Order Form, Survey Form)	Project & Inquiry Based Learning, Demonstration
11th	Introduction to Front-End Framework, Material Bootstrap, Installation, files structure, Syntax & Functionality	Discussion & Concepts
12th	Grid Layout, Typography, icon, Responsive design, navigations, Modal, Card, tables, form.	Discussion, Assignment, Problem Based Learning
13th	Make Mini Project: Simple portfolio pages: Menu, Banner, main content body, contact form, embedded map, footer.	Group Project Based Learning, Group Assignment, Demonstration

C. Web Programming-II (Back-End)

TABLE. III. SECOND WEB COURSE SYLLABUS FOR BACK-END PART

Week	Topics	Strategy & Learning Experience
1st	Web server installations, Configuration, Create Database and PHP files.	Discussion, Troubleshooting & Demonstration
2nd	PHP Basic syntax, commenting, variable, Super Global Variable, loop, array.	Discussion, Assignment, Problem Based Learning
3rd	PHP functions, date time, include, file read write, session, cookies, MySQL connection.	Discussion, Assignment, Problem Based Learning
4th	Basic MySQL syntax, create table, insert, update, delete, joining, design database.	Discussion, Assignment, Problem Based Learning
5th	CRUD: PHP-MySQL form value insertion, multiple value insertion in Database	Problem Based Learning, Demonstration, Assignment
6th	CRUD: PHP-MySQL show lists data from database, searching data, deletion data values.	Problem Based Learning, Demonstration, Assignment
7th	CRUD: PHP-MySQL update database values by selection.	Problem Based Learning, Demonstration, Assignment
8th	Upload single and multiple files, photos in PHP-MySQL	Problem Based Learning, Demonstration, Assignment
9th	PHP-MySQL, Ajax dynamic data insertion, deletion, show.	Problem Based Learning, Demonstration, Assignment
10th	Basic concepts of Model View Controller (MVC), Architecture, example.	Discussion, Overview & Concepts
11th	PHP-MySQL validation, create simple registration and login page and perform insertion and login operation in PHP-MySQL	Project & Inquiry Based Learning, Demonstration
12th	Basic concepts of Framework, Laravel installation, files structure, syntax, architecture.	Discussion, Overview & Concepts
13th	Mini Project: Simple Admin Panel for personal portfolio website.	Group Project Based Learning, Group Assignment, Demonstration

D. Evaluations in CDIO Method

We will follow the CDIO (Conceive, Design, Implement and Operate) which is “learning by doing or project education & learning” [27] to evaluate our proposed syllabus. Because Chen & Fu showed in their paper how CDIO method improved Web development courses significantly [28]. In this method, instructor will plan how to show the teaching content through simple problem-based project. They will examine, design, develop and run the project and student will follow them from the beginning to the end. They may have out-of-class discussion or activity through group assignment. Also, during the session instructor can modify the problem and throw variety of problem to the students. Finally, by developing group project the students will have complete idea of how to use all these chunks of web tool & technological knowledge into combined one and they will develop creativity, teamwork and interpersonal communication skills.

VIII. Conclusion

We have presented an outline for what we believe should form a web engineering curriculum. It may seem a vast syllabus but any instructor can resize the contents based on the categories. These are the minimum view of contents that should appear in any web development programming courses. It contains a simple introduction of front-end and back-end tools & technologies which are mostly required in web related job filed shown in Fig. 1. In order to improve our country's university education, we believe these two course curricula of web engineering can help a lot. Also, there are many scopes of improvement in this proposed a curriculum for web engineering. There may be new contents to be added or removed to adapt latest technologies or we may divide this syllabus into three semester courses to provide enough time to teach and learn.

Acknowledgment

I would like to thank Mr. Mizanur Rahman, Front & Back-end expert developer from Bangladesh University of Business & Technology of his extraordinary support and professional opinion to build these course curricula.

References

[1]. "https://www.daily-sun.com/post/407497/2019/07/15/Digital Bangladesh--a-Story-of-Transformation" The Daily Sun, Web. 15th June 2019.

[2]. Ironman Draft, "ACM/IEEE-CS Joint Task Force on Computing Curricula 2013", ACM-IEEE Society, 2013.

[3]. Xusheng Wang, "A Practical Way to Teach Web Programming in Computer Science". Journal of Computing Sciences in Colleges (2006): 211-220.

[4]. Tuomas Turto, Tommi Mikkonen "A Course on Web Programming". Proceedings of the 1st International Educators' Day on Web Engineering Curricula WECU (2010): 10 p.

[5]. Stepp, M., Miller, J., and Kirst, V. A '"CS 1.5" introduction to web programming', Proceedings of the 40th ACM technical symposium on Computer science education. (2009): 121-125.

[6]. Noonan, R. E. "A course in web programming", Consortium for Computing Sciences in Colleges, Journal of Computing Sciences in Colleges Vol. 22 (2007): 23-28.

[7]. Wang, X., A practical way to teach web programming in computer science, Consortium for Computing Sciences in Colleges, Journal of Computing Sciences in Colleges, Vol. 22 (2006): 211 - 220.

[8]. Xusheng Wang, "Design, Develop and Teach the Second Web Programming Course in Computer Science Curriculum". Journal of Computing Sciences in Colleges (2014): 52-59.

[9]. Lantis, M. J. "Using a web editor as a development platform for teaching HTML and client-side programming in the internet 101 course nifty tools and assignments", Journal of Computing Sciences in Colleges (2008): 97.

[10]. R. W. Connolly "Awakening rip van winkle: modernizing the computer science web curriculum", ITiCSE '11: Proceedings of the 16th annual joint conference on Innovation and technology in computer science education (2011):18–22.

[11]. Robert F. Dugan, "A single semester web programming course model". Journal of Computing Sciences in Colleges (2013): 26-34.

[12]. Francesco Maiorana, "Teaching Web Programming an Approach Rooted in Database Principles". 6th International Conference on Computer Supported Education (2014): 49-56.

[13]. Liu, Y., Phelps, G. "Challenges and professional tools used when teaching web programming". Journal of Computing Sciences in Colleges, Vol. 26 (2011): 116-121.

[14]. Laverty, J. P, "Implementing a dynamic database driven course using LAMP". Information System Education Journal (ISEDJ), Vol. 9 (2011): 33-40.

[15]. Baatard, G. "Teaching PHP with security in mind". In Proceeding of the 5th Australian Information Security Management Conference (2007): 21-27.

[16]. Olan, M., 2009. Web applications: a test bed for advanced topics. Journal of Computing Sciences in Colleges, Vol. 24, No. 3, pp. 72-80.

[17]. Gousie, M. B., "A robust web programming and graphics course for non-majors". ACM SIGCSE Bulletin, Vol. 38 (2006): 72-76.

[18]. Adams, D. R. "Integration early: a new approach to teaching web application development". Journal of Computing Sciences in Colleges, Vol. 23 (2007): 97-104.

[19]. Chao, J., Davey, B. "Navigating the Framework Jungle for Teaching Web Application Development". Issues in Informing Science and Information Technology (2013): 95-109.

[20]. https://roadmap.sh, Developer Roadmaps. Web. 2020.

[21]. http://www.availdata.com/modern-Websites.cfm, AvialData. Web. 2012.

[22]. Verbyla, J., Roberts, G., "Web technology as curriculum". Proceedings of the 3rd Australasian conference on Computer science education (1998).

[23]. Rode, J., "Nonprogrammer web application development". Conference on Human Factors in Computing Systems (2005): 1055 - 1056.

[24]. Yi Liu, Gita Phelps. "Challenges and professional tools used when teaching web programming". Journal of Computing Sciences in College. Vol-26 (2011): 116-121.

[25]. Ironman Draft, "Computer Science Curricula 2013", ACM-IEEE Society, 2013.

[26]. "http://cei.ust.hk/teaching-resources/outcome-based-education/ institutional-resources/obe-principles-and-process". Center for Education Innovation. Web. 2016.

[27]. S. Qiu, "The Application of CDIO Education Mode in the Teaching of Java Language Programming", Modern computer, Vol-9, (2011): 21-24.

[28]. Zhang-bin Chen & Long-tian Fu, "Reform and Exploration of Web-based Training Course Teaching Based on CDIO Mode". 4th International Conference on Education and Social Development (2019).

APPENDIX

TABLE. IV. FINDINGS

University Lists	Web Course	Contents
IIT, University of Dhaka	1	Introduction to Html, Java Script & CSS, Server-Side Programming: HTTP Server, Application Server, MVC Web Framework, Web Services, Database Access: Object Relational Mapping, Lambda Expression, Language Integrated Query, Data Reader, Writer, Web Security: Denial of Service, Buffer Overflow, Cross Site Scripting, Authentication and Access Control
CSE, University of Dhaka	1	N/A
University of Rajshahi	0	N/A
Bangladesh University of Engineering & Technology	0	N/A,
University of Chittagong	0	N/A
Jahangirnagar University	0	N/A
Shahjalal University of Science and Technology	1	N/A
Khulna University	1	Internet and World Wide Web Applications, HTML, SGML, CGI Programming, Active Server Page Programming, Electronic Commerce, Internet Database, Javascript, VB Script, PHP, ASP.NET, Jquery, XML Programming, Flex, WCF, WPF, AJAX, MVC, Silverlight, CMS, Cold Fusion, Python, Mobile web applications.
Hajee Mohammad Danesh Science & Technology University	2	N/A
Mawlana Bhashani Science and Technology University	2	Introduction to the Internet, the web, web 2.0 and Ajax, browser basics, XHTML, cascading style sheets (CSS), JavaScript, Dynamic HTML, XML, RSS, building Ajax-enabled web application, Macromedia Flash, Adobe ® Flex TM , Macromedia ®, Dreamweaver ®, web servers (IIS and Apache), database: SQL, MySQL, DBI and ADO.NET 2.0, web services, PHP, Ruby and Ruby on Rails, ASP>NET, web forms and web controls, JavaServer Pages web applications, Perl and CGI (Common Gateway Interface), etc.
Patuakhali Science and Technology University	1	HTML, CSS, JavaScript, Joomla 2.5 and WordPress, PHP,MySql .It is valuable to both beginners and advanced developers that already have experience in developing web applications.
Noakhali Science and Technology University	N/A	N/A
Rajshahi University of Engineering & Technology	N/A	N/A
Khulna University of Engineering & Technology	1	N/A
Jagannath University	1	N/A
Comilla University	1	Browser and Web Document. Static, Active and Dynamic pages, Programming paradigms and Web programming. Object-oriented vs. Object-based programming, What should and should not be programmed on the Web, Tasks suitable for programming on the Web, Choice of programming language for Web programming. JavaScript for Web Programming: Introduction to the Language, JavaScript: Object Hierarchy and working with objects, JavaScript: Event-Driven Programming, Common Gateway Interface (CGI): Definition, Characteristics, CGI Programming Mechanism: GET and POST methods, Simple examples using Perl, Introduction to PHP Programming Language. PHP for Web Programming
Jessore University of Science & Technology	N/A	N/A
Pabna University of Science and Technology	N/A	N/A
Bangladesh University of Professionals	1	N/A
North South University	1	The course develops an in-depth knowledge of the concepts, principles and implementation techniques related to the Internet and web technology. Details about the Internet, Intranet, Extranet, and e-

		commerce will be covered. Topics include Web server management, threats, security of client and server, network security like firewall, SSL, etc., authentication and authorization, legislation, privacy and IP act, electronic payment, e-business, search engine, Internet protocols like TCP/IP, SGML, XML. Design and development of Web applications using Java Applets, ASP, Java Script, CGI and other Web tools is discussed.
University of Science and Technology Chittagong	1	N/A
Independent University	2	Essential topics such as OSI & TCP/IP architecture, Internet Routing, IP addressing & Domain Name System will be covered. Discussions will be held on popular browsers, HTML and Cascading Style Sheet, HTTP, HTTPS, FTP, Client and Server side scripts, Scripting (JavaScript, AJAX, XML) with jQuery libraries, Web Servers (IIS, Apache). Students will learn to design dynamic websites using ASP.NET with SQL server and PHP with My SQL. A brief overview of topics in web security such as cryptography, digital signatures, digital certificates, authentication & firewall will be provided.
American International University-Bangladesh	1	Introduction and Practical use of HTML & XHTML; Introduction and Practical use of XML; Introduction and Practical use of XSL, & XSLT; use of XQuery, & Schema; XPATH, & XLINK; use of JavaScript; use of PHP; Database Connectivity with PHP; XML use with ASP.NET.
Ahsanullah University of Science and Technology	1	Introduction to Internet technology: Word Wide Web (WWW), Web pages, Web servers, HTTP, HTTPs, FTP, Electronic mail, Search engines, Global databases, digital libraries, video on demand, streaming audio and video; Web page design: HTML and DHTML concepts, tags, commands, form design, table design, online request, dynamic functions, buttons, animations and multimedia, Script languages, Embedding scripts in HTML; Intranet: Usefulness of intranet, Sharing scarce resources over intranet, Network chatting and newsgroups; E-Commerce: Paying money over the network, Online shopping cart, Mobile payment system; Web Security: Privacy Policy, Encryption techniques, Network security and firewalls.
Dhaka International University	0	N/A
East West University	1	Web Fundamentals, Programming Languages for the Web, HTML Basics and the working environment, Fundamentals of PHP language, HTML with PHP, forms, sessions,cookies, etc., CSS and templates, Database manipulation in PHP, Programming the browser and forms withJavaScript, Dynamic programming using Asp.net, AJAX basics,DHTML, Security pitfalls and basic solutions, Lab exercises, Mini project
University of Asia Pacific	1	Introduction to web server and web programming, introduction to any scripting language (such as PHP, JSP), Configuring web server, HTML and Scripting language Tags, Statements and Whitespace web programming, Comments, Functions, Variable Types and Operators, Control Flow, Arrays, HTML forms, Retrieve data from form elements using Get and Post Methods, String Manipulation, Database Connection, Executing SQL queries, Session Control and Cookies, File Handling.
BRAC University	1	A survey of current Internet technologies and state-of-the-art web programming methods. Using client/server structures, topics studied will be drawn from JavaScript, JSP, ASP, Cold Fusion, Flash, Document Object Model, HTML, Cascading Style Sheets, XML, CGI, TCP/IP and the .NET platform. Programming tools may include PERL, various UNIX shell scripts, Windows batch files, Java and other languages as needed.
Manarat International University	1	N/A
Daffodil International University	1	N/A
Green University of Bangladesh	1	N/A
Bangladesh University of Business and Technology	1	Basic design and implementation of websites, Discussion of different navigation and organizational strategies, Client-side technologies including HTML5, CSS, JavaScript, JSON, and JQuery, Server-side technologies emphasizing implementations in PHP, Back-end data management, Interfacing Internet to a database. Querying a database using Cold Fusion, Security issues, Emerging technologies
University of Liberal Arts Bangladesh	1	Designing an Internet utilizing a range of different technologies. Simplifying the creation and updating web content. Expanding Intranet services by adding client-slide and server-side processing. Interfacing Internet to a database. Querying a database using Cold Fusion.

Improving Performance of the Multiplexed Colored QR Codes

Islam M. El-Sheref[1], Fatma A. El-Licy[2], Ahmed H. Asad[3]
Department of Computer Sciences, Faculty of Graduate Studies for Statistical Research
Cairo University, Giza, Egypt

Abstract—**The vast popularity and useful applications of the QR code were the incentives that encourage the research towards improving its storage capacity and performance. A colored quick response (QR) algorithm is proposed, devised and tested to expand the capacity of the black and white modules to hold colored modules that can fold as many as those available in the 8 bits red-Green-Blue (RGB) color code. Fast Multiplexing Technique (FMT) is established to improve the performance and the storage capacity of the QR color code by multiplexing the black/white QR codes into RGB color shades then folding them into the QR code Modules. Comparative experiments with the classical multiplexing technique -MUX system- proved that FMT has a much better performance, (exponentially faster), while maintaining the capacity multitude to 24 folds that of the classical QR code.**

Keywords—Quick response codes; colored quick response codes; data capacity enhancement; multiplexing quick response color coding

I. Introduction

QR stands for Quick Response indicating that the code contents should be decoded very quickly at high speed. It is a 2-dimensional code that have become widely popular in industry, economics, security and education. It is more useful than a standard barcode because it can store (and digitally present) much more data. It has been utilized in diverse applications including Mobile operating systems, Virtual stores, Website login, Wi-Fi network & TOTP, Payment, Robot Navigation and Education. It can hold information and data, with diverse format and types including URL links, Geo coordinates, Wi-Fi and text. QR code, not only supported by smart systems and many of the modern and smart cell phones but also, its encryption feature could be exploit as a security in terms of data tampering [1]. QR codes system became popular, not only, due to its fast readability, but also because of its greater storage capacity [2, 3, 4].

The versions of the QR code range from version 1 to version 40, each of which has different module configuration and number of modules. The module refers to the black and white squares that make up QR code [5, 6, 7]. The capacity of a given QR code version depends upon the error correction level, L (low), M (medium), Q (quartile), H (high) where 7%, 15%, 25% and 30%, respectively, of symbols can be corrected in case the original code has been damaged [7].

The Quick Response (QR) code is invented by the one of major Toyota group companies in 1994 and was initially used for tracking inventory in vehicle parts manufacturing and approved as an ISO international standard (ISO/IEC18004) in June 2000 [8]. A QR code is a type of matrix or two-dimensional barcode that stores up to 4296 characters of information and it is designed to be read by smartphones [4]. A QR code consists of black modules arranged in a square pattern on a white background. The information encoded may be text, a URL or other data [6]. The idea behind the development of the QR code is the limitation of the standard Universal Product Code (UPC) barcode information capacity (that holds, only, 20 characters) [8]. QR codes system became popular due to its fast readability and greater storage capacity [3, 9, 10].

A. Types of Two Dimensional Code

There are many types of 2-D Code [11, 12], each of which, has its unique structure. The structure of the QR code is shown in Fig. 1(A), while, Fig. 1(B) [13], depicts the structure of several other types of two-dimensional codes. Yet, QR codes are utilized most often for its higher data storage capacity. Its specifications offer many more advantages than that of the one-dimensional code. The advantages of QR Code include:

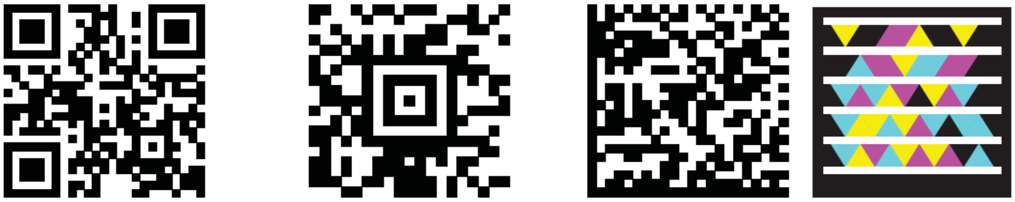
- Reduced space.
- Durability against soil and damage.
- High data capacity.
- Supporting more languages.
- Supporting of 360-degree reading.

B. Improving QR Capabilities

Influenced by QR code structural flexibility, many applications have been focusing in improving QR code data capacity and its tolerance to distortion. Some of these techniques focus on increasing data capacity including data hiding, data compression, Multiplexing and colored QR codes techniques [13- 20].

1) Data hiding techniques: Data hiding techniques distort the QR code during the hiding process, which can be resolved by using additional algorithms for the correction of image distortions. The achieved expansion of the data capacity, however, was limited [14].

2) Data compression techniques: Another way to increase the data capacity is to compress the data before generating a QR code, results, however, shows that data compression technique offers compression up to 52% of the original size [15].



(a) (b)

Fig. 1. QR Code, (B) other Type of 2D Code [13].

The objective of this work is to utilize, colored coding multiplexing for improving the QR code performance and capacity, therefore, the following Section II will discuss the work related to colored QR code and color code Multiplexing. Section III will discuss the proposed system, while Section IV presents the system evaluation through two experiments and illustrates the results. The paper is concluded by the conclusion in Section V and bibliography.

II. Related Work

The research, in both Academia and Industry have been active to improve both the capacity and the tolerance of the QR code. Most of the researches, however, have targeted the QR code capacity without much compensation for its performance. The related research techniques that focus on increasing data capacity includes colored coding [13, 16, 17] and Multiplexing [18, 19, 20].

A. *Colour Coding Techniqiues*

Melgar et al., in [16] purposed a colored QR code structure. It is designed to employ five different RGB colors (red, green, blue, black and white), which enables twice as much as that of the traditional binary QR codes. Each information module represents 2 bits, while the traditional utilize 1 bit for each module. Black modules are used only for alignment purposes. The "red, green, blue and white" colors are chosen because of their maximum equidistance on the RGB color space.

Melgar et al., in [17] proposed CQR Code-9. This code can store up to 2,024 information bits which is twice what was t achieved in their previous research paper [15], and 4 times that of classic QR code. They accomplished this by folding 3 bits of information into the single module.

Blasinki [13], exploit the spectral diversity afforded by the cyan (C), magenta (M), and yellow (Y) print colorant channels, commonly, used for color printing and the complementary red (R), green (G), and blue (B) channels, respectively, used for capturing color images. Specifically, exploiting this spectral diversity to realize a three-fold increase in the data rate by encoding independent data in the C, M, and Y print colorant channels and decoding the data from the complementary R, G, and B channels captured via a mobile phone camera. Developing interference cancellation algorithm to mitigate the effect of cross-channel interference among the print colorant. The authors succeeded, also, in avoiding CMYK/RGB color system interference.

B. *Multiplexing Techniques*

Vongpradhip [18], introduced the Multiplexing QR Codes. His idea was to divide the original data into several portion of smaller size. Each of which formed QR code pattern in its standard form. The generated set of QR codes is multiplexed into special symbols, encoded as black and white modules. The multiplexed QR code is decoded to give back the number of QR code patterns that was multiplexed. The number of required symbols is defined by 2^n, where n is the number of QR code patterns. The authors utilized pattern recognition algorithm and did succeed to increase the data storage capacity; however, it was limited by the number of the symbols that could be recognized.

Gupta [19], purposed new approach that doubled the storage and accelerated QR decoding. The author divides the data into two portions, then converted each into a classical QR code. Where He encoded the data in, only, one third of the module. Yet, he rotated the second QR code 90 degrees before encoding, then placed it on the first QR code (to form vertical, horizontal and cross shaped black modules). The author performed the decoding by utilizing bar code decoding system.

Galiyawala [20], introduced multiplexing (MUX) method, in which several QR codes are multiplexed to generate a single colored QR code. This technique offered improvement of the data storage capacity up to 24 times that of QR code of the same version. He divided original information into 'n' portions, each was encoded into a classic QR code. That is, to generate 'n' individual QR codes which needed 2^n distinct encoding colors. The author multiplexed the individual set of n-bit (a corresponding bit from each of the generated n QR codes) to the corresponding color from a look up table for encoding. In decoding, however, each color shade was demultiplexed by accessing its corresponding code from the look up table, to obtain n-bit code, representing the corresponding bit in each of the n classic QR codes. The author employed a table for the color coding to be accessed for encoding the 'n' multiplexed bits into the corresponding color. The colored QR code was demultiplexed by searching the table for the color's code to generate the original 'n' QR codes. Galiyawala's procedure, therefore, needed n-entry color-coding table for every given n.

III. Proposed System: Fast Multiplexing Technique

The main objective of the proposed Fast Multiplexing Technique (FMT) is accelerating the multiplexing methodology of Galiyawala [20]. This is accomplished by utilizing the natural Red-Green-Blue (RGB) color coding as three-Dimensional array. RGB is adopted to utilize its maximum equidistance of the Red-Green-Blue color shades. Each single point in this three-dimension space represents a unique code of a color shade. Given that each of the three RGB has 256 different shades, the maximum number of color shades is 256^3. In order to represent n different colors, one need 2^n binary digits. The resolution of the color shade coded into the colored QR code depends upon the data size and the QR code version.

The communicated data is divided into several portions, each of which, is encoded into a classical QR code that represent a layer in the multiplexing QR codes as shown in Fig. 2. The black module in the QR code is, normally, represented as ‘1’, while the white module is represented as ‘0’. The binary value obtained from the corresponding bits (modules) on each layer (as illustrated by the vertical rectangular in Fig. 2(B)) is encoded to the unique color shade of the RGB colors and folded into a single module. The maximum number ‘n’ of possible portions/layers, is:

$$n= \log_2(256^3)= \log_2((2^8){}^\wedge 3)= 8\times 3= 24 \text{ portions/layers} \qquad (1)$$

The integrity of the color is guaranteed by including number of layers as the folding length in a predefined set of modules to be referenced in the decoding stage. The decoding process is carried out in the reversed steps: decoding colors, generating the several black/white layers of QR codes and decoding each into its original portions accordingly.

A. Encoding Process

Given a data of size M, the function Info-multiplexing calculates the appropriate number of portions/layers ‘L’, (depending on the version’s capacity of the Classic QR image). Accordingly, the required number of color codes is 2^L. The RGB color resolution is calculated so that the shades are distributed as evenly as possible. That is, ‘L’ is divided three ways, so that each of the RGB colors has at least $2^{L/3}$ different shades. Assuming that L was divisible by 3, then the values corresponding to each of the Red-Green-Blue colors ranges from 0 to $2^{L/3}$. This range defines the number of shades per color, SPC (CPR, SPG and SPB respectively). The string of L binary bits obtained from the corresponding black/white modules is denoted by bit-string. The bit-string is converted to the corresponding resolution of the color palette of the SPCs. The following are the steps of Info-Multiplexing procedure:

1) Calculate Number of data portions, ‘L’;

2) Divide data into L QR Codes;

3) Initiate the new colored QR code by encoding the value of ‘L’ in the first set of data modules in black and white;

4) Calculate the SPCs for the three R-G-B colors;

5) For each bit-string of data modules do;

a) Convert its binary value into decimal value ‘D’;

b) Calculate the resolution of ‘D’ in the SPCs palette, ‘M’;

c) Get the corresponding of ‘M’ in RGB color system, ‘C’;

d) Print colored module, ‘C’;

B. Decoding Process

Given a colored QR code, the procedure Color-demultiplexing employs the parameters in the colored QR code to retrieve the number of folds in each module, accordingly, it calculates the color resolution in RGB and decode each color/module into its corresponding binary value. It generates the set of black/white QR codes, and the corresponding portions of the data set. The following are the steps of Color-Demultiplexing procedure:

1) Read the number of layers ‘L’;

2) Calculate the SPCs for the three R-G-B colors;

3) For each colored module ‘C’ in the colored QR code do:

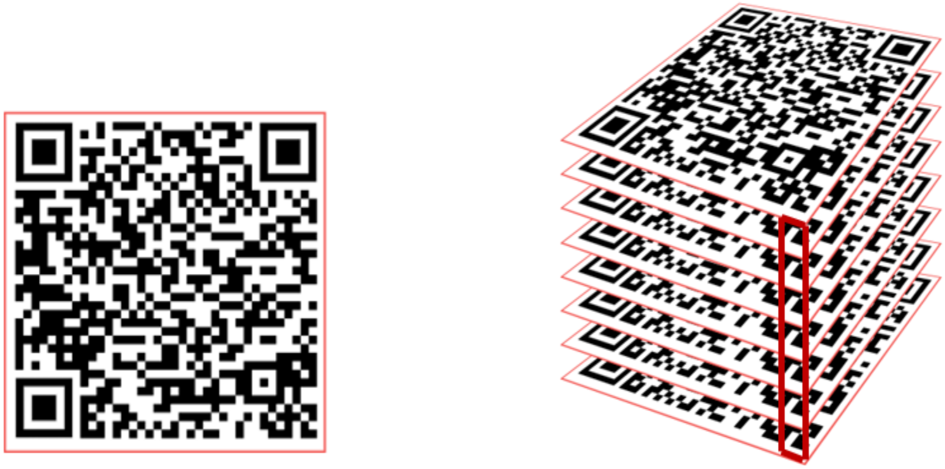
a) Calculate the resolution of the shade of ‘C’, ‘M’;

b) Calculate the decimal value ‘D’ of ‘M’;

c) Calculate the binary value of D;

d) Generate the corresponding ‘L’ black and white modules;

4) Decode the L black/white QR codes into its original data.



(a) (b)

Fig. 2. (A) Classic QR Code. (B) 8 Layers of QR Codes.

IV. Evaluation

The proposed system was tested and then evaluated through two experiments. The first was a comparative study of the system performance against that of Galiyawala's classical multiplexing system [20]. The second experiment was performed to measure the capabilities of the system, in handling bigger data size, by utilizing images and picture instead of text files.

All the experiments were performed in server scripting language (php)[1], and are carried out on a computer with processor specification: Intel(R) Core (TM) i7-3770M CPU @ 3.40GHz and 64-bit windows 10 Pro.

A. *Comparison with Mux Technique*

The first experiment is performed to test the FMT system performance against that of Galiyawala [20]. Their MUX experiments were performed in MATLAB and carried out on a computer Intel(R) Core (TM) i3-2310M, CPU @ 2.10GHz and 64-bit windows 7 Ultimate operating system. QR code of version 2 (has 252 modules hold maximum of 255 character of information) is used for their experiment. For the sake of comparison, sets of randomly generated data are prepared to occupy the same number of the QR codes as those utilized in [20], each, are applied as input to the FMT system to perform the encode/decode processes for the corresponding colored QR Codes. The system assumed version number 40-L (has 1772 modules, with low level error correction), instead of version 2 to compensate for the faster processor and the current operating system employed for FMT. Table I enlists all the observations and results of the FMT experiments. Column one represents number of QR codes multiplexed, which corresponds to the number of multiplexed layers L. Fig. 3 illustrates a plotting of encoding/decoding execution time versus the number of multiplexed layers for 13 different sets of data (shown in Table I). This figure includes, also, the plotting of the results for encoding and decoding observations of the MUX system [20], which have been scaled down.

The plots are built with the processing time units in milliseconds, yet, the results of the MUX system [20] are scaled down to the $\log_{10}$ of its original values, due to the exponential differences between them and those of the FMT for the corresponding data sets. For 8 Layers of QR codes, FMT consumes 0.383 and 0.226 milliseconds for encoding and decoding, respectively, while, MUX consumes 3911 milliseconds for encoding and 4359 milliseconds for decoding. MUX consumes much more time for decoding than encoding, especially when dealing with more than 9 layers of QR codes.

B. *Measuring FMT Capabilities*

A second experiment is performed to measure the performance of FMT in processing several different images' sizes. A set of different size images/pictures are applied as real information. FMT is exercised to multiplex each of the images into colored QR code, then demultiplex the colored code back to its originality. The observations and results for encoding and decoding of 11 different size pictures is depicted in Fig. 4. The proposed system can handle images of size up to 51 KB. The time for decoding is, in general, less than that for encoding, for the 51 KB image it consumes less than 0.9 milliseconds.

TABLE. I. Observations and Results of FMT System Multiplexing Number of QR Codes

Number of QR codes multiplexed	**Assigned distinct colors (2^n)**	**Character size (Binary) 40-L**	**Encoding time (millisecond)**	**Decoding time (millisecond)**
2	4	5906	0	0
3	8	8859	0.198	0.188
4	16	11812	0.205	0.203
5	32	14765	0.212	0.209
6	64	17718	0.213	0.211
7	128	20671	0.265	0.215
8	265	23624	0.383	0.226
9	512	26577	0.389	0.298
10	1024	29530	0.393	0.32
11	2048	32483	0.396	0.359
12	4096	35436	0.4	0.389
13	8192	38389	0.411	0.396
14	16384	41342	0.423	0.412

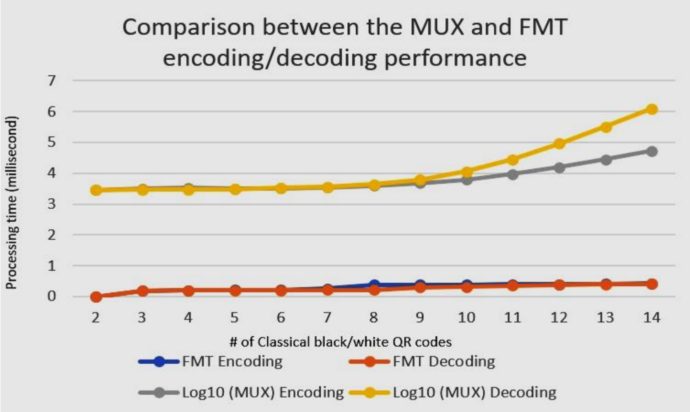


Fig. 3. Performance of the FMT System, Versus that of MUX.

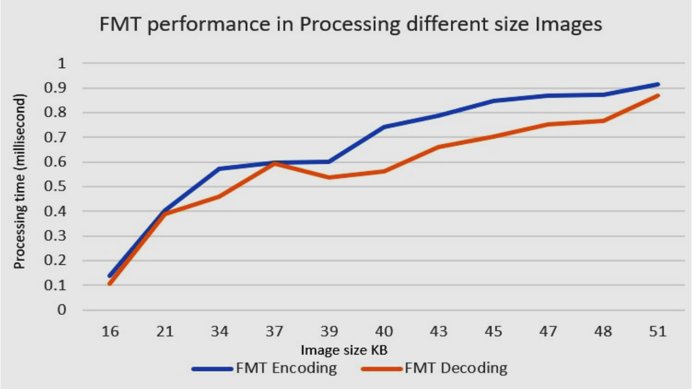


Fig. 4. FMT Performance in processing different Size's Images.

[1] https://igy-apps.com/fmt/

V. Conclusion

The presented technique has a firm ground for improving the colored QR performance and capacity. It provided a possibility of encoding different kind of data with as much size as it could be hold in 24 folds of the given QR version. The technique based on utilizing RGB color scheme extension as a third dimension of the QR code. The FMT system proved to be very efficient, as it is exponentially faster than the classical MUX technique, yet, possessing the same property of multitude the capacity of the QR code. Its processing time depends, only, on the number of multiplexed QR codes. The FMT technique improved the capacity of QR code up to 24th fold, while preserving its intrinsic property of having Quick Response.

The presented technique did not consider the problems associated with colors distortions caused by printer while printing QR code or by camera sensor and surrounding lighting during the capture process. Yet, if the means of the data transmission is purely digital, these problems can be disregarded.

Future utilization of the proposed technique is to integrate it into a QR scanning system to provide a colored coded QR version with the merit of much bigger capacity and better performance.

References

[1] S. Tiwari and S. Sahu, A novel approach for the detection of OMR sheet tampering using encrypted QR code, in: IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, 2014, pp. 1-5.

[2] N. Bhardwaj, R. Kumar, R. Verma, A. Jindal and AP. Bhondekar, Decoding algorithm for color QR code: A mobile scanner application, in: International Conf. on Recent Trends in Information Technology (ICRTIT), India, 2016, pp. 1-6.

[3] S. Goyal S, S. Yadav and M. Mathuria, Exploring concept of QR code and its benefits in digital education system, Advances in Computer Science and Information Technology (ACSIT), 3(5) (2016) 452-456.

[4] S. Tiwari, An introduction to QR code technology, in: 2016 IEEE International Conference on Information Technology (ICIT), Bhubaneswar, 2016, pp. 39-44.

[5] S. Ahlawat, C. Rana and R. Sindhu, A review on QR codes: colored and image embedded, International Journal of Advanced Research in Computer Science, (IJARCS), 8(5) (2017) 410-413.

[6] A. Singh and P. Singh, A review: QR codes and its image pre-processing method, International Journal of Science, Engineering and Technology Research (IJSETR), 5(6) (2016) 1955-1960.

[7] K. H. Pandya and H. J. Galiyawala, A Survey on QR codes: In Context of Research and Application, International Journal of Emerging Technology and Advanced Engineering (IJETAE), 4(3) (2014) 258-262.

[8] T. J. Soon, Section three the QR Code, The Synthesis Journal: iTSC Information Thechnology Standard Comitee, Singapore, (2008) 59–78.

[9] Z. Čović, Ü. Viktor, J. Simon, D. Dobrilović, Ž. Stojanov, Usage of QR Codes in Web Based System for the Electronic Market Research, in: 14th IEEE International Symposium on Intelligent Systems and Informatics (SISY), Serbia, 2016, pp. 187-192.

[10] R. I. Rizqi, N. A. Rohama and DR. K. Nimkerdphol, Inventory management system using QR code on android a case Study in computer engineering department, Journal of Electrical Engineering and Computer Sciences (JEECS), 3(1) (2018) 381-388.

[11] A. Grillo, A. Lentini, M. Querini and G. F. Italiano, High Capacity Colored Two Dimensional Codes, in: Proceedings of the International Multiconference on Computer Science and Information Technology, 2010, pp. 709–716.

[12] H. Bagherinia and R. Manduchi, A novel approach for color barcode decoding using smart phones, in: IEEE International Conference on Image Processing (ICIP), IEEE, 2014, pp. 2556–2559.

[13] H. Blasinski, Per-Colorant-Channel Color Barcodes for Mobile Applications: An Interference Cancellation Framework, IEEE Transactions on Image Processing, 22(4) (2013) 1498-1511.

[14] R. Paul, A review on Stegnography in QR codes, International Research Journal of Engineering and Technology (IRJET), 5(5) (2018) 4294-4296.

[15] A. Abas, Y. Yusof and F. Kabir, Expanding the data capacity of QR codes using multiple compression algorithms and base64 (encode/decode), Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 9(2-2) (2017) 41-47.

[16] M. E. Melgar, A. Zaghetto, B. Macchiavello and A. Nascimento, CQR codes: Colored quick-response codes, in: IEEE 2nd Inter. Conf. on Consumer Electronics, Berlin (ICCE-Berlin), 2012, pp. 3321-325.

[17] M. E. Melgar, M. Farias, F. Vidal and A. Zaghetto, A High Density Colored 2D-Barcode: CQR Code-9, in: 29th SIBGRAPI Conf. on Graphics, Patterns and Images, Sao Jose dos Campos, 2016, pp. 329-334.

[18] S. Vongpradhip, Use Multiplexing to Increase Information in QR Code, in: 8th International Conf. on Computer Science & Education (ICCSE), Colombo, Sri Lanka, 2013, pp.361-364.

[19] K. D. Gupta, M. Ahsan and S. Andrei, Extending the storage capacity and noise reduction of a faster QR-code, Broad Research in Artificial Intelligence and Neuroscience (BRAIN), 9(1) (2018) 59-71.

[20] H. J. Galiyawala and K. H. Pandya, To Increase Data Capacity of QR Code Using Multiplexing with Color Coding: An example of Embedding Speech Signal in QR Code, in: IEEE India Conference (INDICON), 2014, pp. 1-6.

An Enhanced Twitter Corpus for the Classification of Arabic Speech Acts

Majdi Ahed[1], Bassam H. Hammo[2], Mohammad A. M. Abushariah[3]
Department of Computer Science[1]
Department of Computer Information Systems[2, 3]
King Abdullah II School of Information Technology, The University of Jordan, Amman, Jordan[1, 2, 3]

Abstract—**Twitter has gained wide attention as a major social media platform where many topics are discussed on daily basis through millions of tweets. A tweet can be viewed as a speech act (SA), which is an utterance for presenting information, hiding indirect meaning, or carrying out an action. According to SA theory, SA can represent an assertion, a question, a recommendation, or many other things. In this paper, we tackle the problem of constructing a reference corpus of Arabic tweets for the classification of Arabic speech acts. We refer to this corpus as the Arabic Tweets Speech Act Corpus (ArTSAC). It is an enhancement of a modern standard Arabic (MSA) tweet corpus of speech acts called ArSAS. ArTSAC is more advantageous than ArSAS in terms of its richness of annotated features. The goal of ArTSAC is twofold: Firstly, to understand the purpose and intention of tweets which act in accordance with the SA theory, and hence positively influencing the development of many natural language processing (NLP) applications. Secondly, as a future goal, to be used as a benchmark annotated dataset for testing and evaluating state-of-the-art Arabic SA classification algorithms and applications. ArTSAC has been put in practice to classify Arabic tweets containing speech acts using the Support Vector Machine (SVM) classification algorithm. The results of the experiments show that the enhanced ArTSAC corpus achieved an average precision of 90.6% and an F-score of 89.6%. Substantially it outperformed the results of its predecessor ArTSAC corpus.**

Keywords—Arabic speech acts; twitter; modern standard Arabic; speech act classification

I. Introduction

People discuss different issues and topics on twitter throughout their tweets. Recently, twitter has gained great attention and attraction from the popular press and, increasingly, from scholars. Speech Act (SA) is an utterance (i.e. a spoken word, statement, or vocal sound) that can be used to present information and also to carry out actions. The idea of a SA can be captured by emphasizing that "by saying something, we do something" [1]. For example, when you ask someone to do something in a sentence like: "Please be quiet"; your utterance represents a request SA. Speech Act Classification (SAC) is the task by which a certain utterance is assigned to a certain predefined SA label such as: assertion, request, etc. based on the content of that utterance. SAC is a traditional classification problem similar to the problem of text classification. Topics that are usually discussed on tweeter represent the subjects of the tweets. These topics are classified into three main types: [2].

1) Entity-oriented topics: Topics about different entities such as famous people (e.g. King Hussein of Jordan), or famous restaurants (e.g. Pizza Hut).

2) Event-oriented topics: Topics about different events and occasions around the world. They are usually about breaking news (e.g. parliament elections in Jordan).

3) Long-standing topics: Topics that are continually discussed on twitter, such as weather, movies, or sports.

Speech Act Theory (SAT) is a linguistic theory that was introduced to formalize speakers' intentions and put them into perspective [3]. SAT aims to understand the utterance defined in terms of a speaker's intention and the effect it has on a listener.

Twitter is one of the big-data sources found on social media. It has hundreds of millions of users who generate around 500 million tweets per day [4]. Due to the tremendous volume of tweets, the problem of classifying and extracting useful information out of them is actually a sort of managing big data. This task can be viewed as a major concern to the field of Data Mining (DM). DM uses different approaches such as classification, association rules, or clustering techniques to discover knowledge in big data.

Defining a catalog of labels (classes) and predicting the label of any given instance based on this catalog is the main goal of classification algorithms. Training a computer machine to classify and label speakers' intentions (retrieved from their utterances) could be viewed as a traditional classification problem. For the problem we are attempting to solve in this study, a catalog of speakers' intentions such as requests, questions, promises, threats, etc. is defined. Then, a classification algorithm is used to discover the speaker's utterance. Such automated utterance classification could be handy in tasks like polarity or sentiment analysis of speakers on social media.

Tweets are usually delivered in a natural language. This fact shows that one of the joint research fields that are heavily indulged in the phenomena of big data is Natural Language Processing (NLP). Generally speaking, a tweet is a short text that usually conveys a single SA. SA classifiers can be used as an initial phase in many contextual mining and NLP tasks such as sentiment analysis, opinion mining, question answering, and rumor detection.

For example, in the case of rumor detection on a social media platform such as Twitter; a SA classifier is needed to classify different tweets and select the ones that might have rumors. Due to the fact that tweets are microblogs (traditionally 140-character per tweet), they make a good source for SAs classification.

Discovering the SA of tweets could also be used in various NLP tasks such as customer polarity. For instance; assume a company wants to measure the degree of satisfaction of its customers about a certain product such as a new mobile phone. Posts on such a product could be in tens or hundreds of thousands. Manual measurement of customer satisfaction in such situations is very hard and time-consuming; hence the existence of an automated approach to accomplish this task could be very helpful.

Arabic is the fifth widely used language in the world. It is the native language of more than 400 million people. Arabic scripts come in three forms: Classical Arabic; like the holy Quran verses, Modern Standard Arabic (MSA) such as everyday formal press statements or news announcements, and Colloquial Arabic like the native dialect of different Arabic countries [5]. In this paper, we are focusing on MSA language which is a formal language that is understood across all Arabic countries. MSA is a light form of classical Arabic that uses only a well-known and common vocabulary. It maintains a formal but simple and easy-going form. Although classification of speech acts is an active research area for the English language [6, 7, 8]; however, there seems to be a little work done on similar research for Arabic language [9].

The importance of this study is driven by the following facts:

1) SA classification can be used to understand the purpose and the intention behind people's tweets. Knowing the SA behind a tweet could allow us to comprehend the mental and emotional state of the tweeters. Predicting the type of SA of tweets about a certain topic can reveal a lot about people's perspectives or attitudes about that topic. For example, a lot of tweets asking about a certain topic reveal that people are confused about that topic or they are mad and demanding actions about it.

2) Classification features for the English language may not be the same for the Arabic language. It is known that different languages rely on different syntax and semantic characteristics to extract the SAC features. This does not eliminate the fact that some features, such as the question mark at the end of a sentence, represent a cross-lingual extraction feature that classifies such a sentence as a question regardless of the language of the sentence (i.e. universality of SA theory).

3) NLP tasks such as sentiment analysis [10], rumor detection [11], and evaluation of customer satisfaction are important in many online applications today; especially in big data environments where the need for automated tools is urgent.

4) NLP research oriented towards Arabic text is limited [12, 13] and, hence there is a dire need for general purpose Arabic language pre-processing tools and benchmark annotated corpora. The proposed classifier and annotated corpus could be of good value in this regard.

In this paper, we tackle the problem of creating a reference corpus of Arabic tweets for the classification of Arabic speech acts. We refer to this corpus as the Arabic Tweets Speech Act Corpus (ArTSAC). It is an enhancement of a modern standard Arabic (MSA) tweet corpus of speech acts called ArSAS [33]. ArTSAC is more advantageous than ArSAS in terms of its richness of annotated features. The goal of ArTSAC is twofold: Firstly, to understand the purpose and intention of people's tweets which comply with the SA theory, and hence positively influencing the development of many Arabic NLP applications. Secondly, as a future goal, to be used a benchmark annotated dataset for testing and evaluating many Arabic SA classification algorithms and applications.

The remaining of the paper is organized as follows: Section II presents the related work to be used to develop a solution for the aforementioned problem. Section III gives a detailed description of the modified corpus. Section IV discusses the development of the classifier and provides an evaluation of its results. Finally, Section V concludes the work and draws a roadmap for the future work.

II. Related Works

We will limit our literature review to automated SA classifiers developed for English and Arabic languages. Many automated SA classifiers for the English language exist, some are dedicated to Twitter. The earliest attempts to build automated classifiers were oriented towards emails.

An SA classifier for emails and Internet forums was presented in [14]. The authors aimed to use the SAs of an email to identify the intentions of its sender. For example, a simple reply in the e-mail's subject field could indicate a reply to a previous request or a question.

In [15] the authors also worked on emails SAs. They demonstrated that the contextual features of an email can improve that email's SAC. In other words, the syntax and semantic features of an email's text can be used to classify an email. The concept of ontology to classify emails according to the sender's intention was introduced by [16]. The proposed ontology consisted of nouns and verbs that could indicate certain intentions. Applying the ontology produced good results for some nouns and verbs. One drawback of this study was the small size of the proposed ontology and its limitation to simple nouns and verbs.

In [17] the authors developed an annotated SA classifier for the classification of online German discussions. They used an n-grams approach to extract the features. The authors achieved better results with similar previous work. An online chat SA classifier was introduced in [18]. In this work, the authors argued that the first few words in each chat were very predictive of its SA category. They believed that the hearer usually infers the speaker's intention after hearing only a few words of the speaker's utterance. For example, a polite request utterance usually contains the word "please" among its first few words. However, we believe that the works of [17] and [18] neglected the role of discourse and speakers' expectations which are very important in an online chat system. In other words, the expected SA of an utterance is affected by the SA of its previous utterance in a conversation. Hence, it is obvious that online chats resemble conversations with a discourse. For instance, the expectation after someone greets someone else is to hear a greeting reply. Similarly, after a question, an answer is expected.

An automated SA classifier for educational games was introduced in [19]. The authors argued that the SA taxonomy should be established by using subject matter experts. They believed that a small set of well-defined SA categories were better than many sophisticated categories and that balanced data sets could be misleading. Also, they argued that the data set should be tailored according to real-world applications because the real data set that a classifier may run on in the future may be unbalanced. Their experiments showed no conclusive results for their last assumption regarding data set the balance.

The work of [20] brought attention to Twitter. SA recognition from tweets is considered a classification task. Thus, the primary work was to find a set of robust features appropriate for solid classification. They argued that SAs provide good insights into the communicative behavior of tweeters on Twitter. Again, one of the problems found in this paper is the lack of a benchmark annotated data set as the authors labeled and used their own tweets. The work of [21] was a continuation of their previous work described in [20]. In this work, the authors enhanced the annotated data set and used different classification algorithms than the ones they used earlier for the purpose of comparison.

A new SA classifier was developed by [22]. A new annotated dataset of tweets was processed and constructed. In this study, an enormous number of features (nearly 2000) were extracted and processed. What made this possible was the availability of many pre-processing tools that helped in automatically defining and extracting those features.

In [23], the author proposed a SA analysis of celebrities' tweets. The study showed that celebrities talked to different audiences using different SAs. In his study, the author used the CMC SA taxonomy, which contains 16 categories of SAs [24]. However, such a fine categorization could be problematic for the classifier. The author reported that few SAs did not appear in any tweet.

An automated jihadist messages' detector for twitter was introduced in [25]. In this work, no manual annotation was used. This was because radical tweets used to train the classifier were taken from known jihadist accounts, and those tweets were presumed radical based on their radical tweeters. This form of assumption could tailor or overfit the classifier for certain features. These features could be person stylistic or not broad enough to generalize. We believe so because the result obtained by the classifier were remarkably high (from 89% up to 100%) depending on the dataset. This does not agree with the modest result obtained by other research discussed in our review.

With respect to Arabic SA classification, [26] pointed out that the work in this field is very humble. Here we present a few related studies. In [27], the authors reported on an experimental study of manual annotation of around 400 newspaper sentences. They were processed using two classification algorithms to produce an SA classifier. In their work, they used techniques such as part-of-speech tagging, named entities, and utterance initial words. What was noticeable in this work that the size of the dataset was very small, and the dataset was not representative; some SA classes have many more instances when compared to other classes, so the data set was considered unbalanced. In addition, a single annotator was used in the experiment, usually, more annotators are needed, and an annotation policy should be used.

Another simplified Arabic SA classifier had been described in the studies of [28] and [29]. In their work, the classifier only focused on classifying questions and non-questions utterances. The classifier was used in a conversational agent called ArabChat in order for the agent to determine questions and answer them appropriately. The proposed agent processed the user's utterances through pattern matching and compared them to predefined patterns which represent different topics.

Many research works based on manual non-automated SAs classifications for classical Arabic scripts had been described in [30, 31, 32]. It was argued in these studies that certain SA frequencies may increase depending on the communicative nature of the discourse under study. Hence, SAs classifications cannot be performed in a complete context-free manner without taking into consideration the situation in which the speaker uttered his words.

In [33], the Arabic SA and Sentiment corpus (ArSAS) was described. The corpus contained a set of around (21,000) MSA tweets. Each tweet in the corpus was annotated with an SA label and a confidence factor of annotation for that label. The availability of a specialized corpus such as the ArSAS can highly advance the research in Arabic SAs. The work of [34] is such an example. In this work, the authors developed an Arabic SA classifier for Arabic tweets using both SVM and deep learning algorithms.

From the previous studies we could derive the following conclusions:

1) Researchers are still following the SA taxonomy described in [38]. There was a little variation to tailor the SA taxonomy.

2) There is plenty of room for improvement in SA feature extraction; consequently, an improvement in SA classifications.

3) A benchmark SA corpus to be used across the field is in high demand.

This research modifies the ArSAS corpus of [33] and the work of [34]. The newly constructed Arabic tweets SA corpus (ArTSAC) is richer than ArSAS in terms of introducing new annotations. ArTSAC will be used to train a classification algorithm to classify Arabic SA tweets according to the SA theory and to be used as a benchmark annotated dataset for testing and evaluating many Arabic SA classification algorithms.

III. Proposed Arabic Tweets Act Classifier

Our goal is to create an Arabic SA corpus of tweets rich in annotated features that can be used in classifying SAs, sentiment polarity, sentiment mining, and other NLP applications. Classification of SAs requires two major components: (1) a reference SA annotated corpus and (2) a suitable classifier. In this section, we discuss in detail the construction of the Arabic Tweets Speech Act reference corpus (ArTSAC) for MSA. Next, we present the Support Vector Machine (SVM) classifier to be used throughout the experiments conducted on the corpus to classify Arabic SA tweets.

A. *Construction of the ArTSAC Reference Corpus*

The construction of an annotated corpus is an essential step to develop any SA recognition system. The construction of such a corpus is a labor-intensive task. Our proposed ArTSAC corpus for modern standard Arabic SA tweets is a modified version of an open-source SA corpus named ArSAS. The construction of ArTSAC required collecting the Arabic tweets from the ArSAS corpus, extracting all their features and annotating them properly, compiling the list of features, and generating the coded file for the classifier. Fig. 1 illustrates the flow diagram for the constructing the ArTSAC reference corpus. The below subsections are the detailed description of each step towards building the corpus.

1) The collection of arabic tweets: Arabic tweets were obtained from an open-source corpus named ArSAS [33]. It has been developed to experiment with Arabic speech acts and it contains about (21,000) MSA tweets. The tweets of ArSAS were classified according to the SA taxonomy described in [20] and they were organized into one of the following classes/categories:

- Assertions: for example, “سيارتي أسرع من سيارتك” (My car is faster than yours.) It indicates that the speaker commits himself to the truth of what he uttered.
- Expressions: for example, “لقد حزنت لما حدث لسيارتي” (I was sad for what happened to my car.) It indicates an expression of emotion by the speaker.
- Requests: for example, “هل تساعدني في تنظيف سيارتي؟” (Can you help me clean my car?) It indicates a request for service or help made by the speaker.
- Questions: for example, “هل تعلم أين مفتاح سيارتي؟” (Do you know where my car’s key is?) It indicates an inquiry about information made by the speaker.
- Recommendations: for example, “يجب أن تستشير الطبيب” (You have to see a doctor.) It indicates advice or recommendation presented by the speaker.
- Miscellaneous: They include different SAs. However, they have relatively few occurrences on Twitter, not enough to warrant a separate category.

A one-to-one association between each tweet and one of the SAs categories was already maintained in the ArSAS corpus. We were careful to make sure that each SA category has enough instances (i.e. tweets) to allow us to robustly define their features. Table I lists the number of tweets for each SA category.

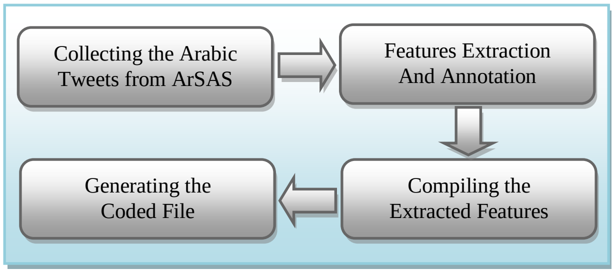

Collecting the Arabic Tweets from ArSAS
Features Extraction And Annotation
Compiling the Extracted Features
Generating the Coded File

Fig. 1. The Flow Diagram for the Construction of the ArTSAC Corpus.

TABLE. I. THE NUMBER OF TWEETS FOR EACH SA CATEGORY (FROM THE ARSAS CORPUS [33])

Tweet’s Class	**Number of Tweets**
Assertion	8203
Expression	11689
Request	183
Question	749
Recommendation	107
Miscellaneous	60
Total	**20991**

2) Features extraction and annotation: The second step in the development process of the ArTSAC corpus was to extract all proper features from the tweets. Features are the pieces of information or properties within the tweets’ messages that convey speech acts. These features represent what a classifier is looking for in order to classify a tweet into one of the aforementioned SA categories listed in Table I.

In order to have the proper guidelines in the feature extraction process of SAs, we conducted a manual analysis of the ArSAS tweets [33] to make sure we have solid insight into the analysis of SAs and their required features.

To conduct the feature extraction task, we got help from two annotators. After we explain to them how to do the features extraction by examples, we performed a pilot task to ensure they understood how to carry out the task. Finally, we could define and extract the following features:

a) Keywords: Some words in a tweet convey certain SA messages. For instance, in an utterance such as “ هل بإمكانك مساعدتي رجاءً” (*could you help me please*), the word “رجاءً” (*please*) usually indicates a request SA. This process was an intensive manual process where we have asked the participants to extract up to eight keywords from each tweet in the ArSAS corpus. After we explained to the participants what they have to do, each participant has produced his own list of keywords. After we aggregated the two lists into one featured keywords list, we obtained 1656 unique keywords. The keywords list mainly included constructs such as proper names and nouns.

b) Twitter special characters: Special characters in tweets might designate certain SAs. For example, a special character that is widely used in Twitter is the hashtag ‘#’. Usually, it indicates an assertion SA. We extracted these special characters automatically using their Unicode values.

c) Topic Label: Each tweet in the ArSAS corpus already has been annotated with a topic label. Labels include Entity, Event, and Long-Standing topics. To the best of our knowledge, this feature was never attempted before in the classification of SAs.

d) Punctuation marks: Few punctuation marks indicate certain SA categories. For example, the question mark ‘?’ usually signals a question or a request SA. Punctuation marks were extracted automatically from the ArSAS tweets corpus.

e) N-grams: Basically, a textual *n*-gram is a sequence of contiguous *n* words that usually co-occur together. N-grams are commonly used in many NLP applications and they usually can help in conveying certain SA messages. For example, the phrase “ألا تعتقد” (*Do you think*) usually indicates a question SA. To extract *n*-grams from the tweets in the ArSAS corpus, we perform manual *n*-gram selection with the help of the participants. Each annotator has selected up to 6 possible *n*-grams for each tweet. No limitation was applied to the size of the *n*-gram segments as many *n*-grams represent verses from the holy Quran, popular quotes, or idioms that could span the entire content of some tweets. However, most of the extracted *n*-gram features were *bi*-gram and *tri*-gram. Other possible segments were 4-grams and 5-grams. Finally, the compiled lists of the annotators have been aggregated into one list of 2658 unique *n*-gram features.

f) Emoticons: Expressing emotions through icons are widely used in social media. Emoticons expressing happiness, sadness, etc. are highly informative in reflecting tweeters' attitudes and moods; hence they can convey certain types of SAs. We automatically extracted emoticons from the ArSAS tweets and compiled a list of 68 emoticons.

g) Links: Hyperlinks are impeded in many tweets. They point to different locations and they possibly could indicate certain types of SAs. Hyperlinks have defined structure, which made extracting them automatically an easy task.

h) Sentiment label: Every tweet in the ArSAS corpus had been already annotated with a sentiment label (positive, negative, mixed, or neutral). We used the sentiment features in the classification process as they may convey certain SAs such as recommendations or assertions. Up to our knowledge and from the literature, the sentiment features have never been attempted in the classification problem of SAs.

i) Tweet's length: Tweets are varying in length. Usually, there is a correlation between the tweet's length and the SA within the body of the tweet. Our analysis of ArSAS showed, for instance, that an expression tweet is usually longer in size than a request tweet. Accordingly, for this feature, we assumed that a long tweet is one that has more than 50 characters; otherwise, it is considered a short tweet.

At the end of the feature extraction task, we could draw the following conclusions:

- Feature extraction was performed automatically and manually. The automatic task was the easiest. It has been applied to extract well-defined features such as special characters and emoticons. For automatic annotation, a set of tools was developed. Each tool was used to extract specific features as discussed earlier. The following pseudo-code is a generalized form of the algorithm LookupTableConstructor. This table is accessed by all tools to construct the ArTSAC corpus. Table II shows a sample of the generated features in the LookupTable.

```
Algorithm: LookUpTableConstructor( )
Pre-request: features
Process:
    while there are more tweets
      read a tweet's feature from ArSAS
      if the feature is not null, then
        search for the feature in the feature's LookupTable
      if not exist, then
        add a feature to the last location in feature's
        LookupTable
    end if
    end if
    end while
 Results: feature's LookupTable
```

- The manual feature extraction task was conducted by annotators through processing 21,000 tweets from the ArSAS corpus. Although the manual analysis was an intensive task, it was essential to get an in-depth understanding of the characteristics and different usages of SAs. The manual process was used to extract five features.

- Only annotations that have been agreed upon by both annotators have been aggregated and included in the features lists of our ArTSAC corpus. Table III shows a summarization of the extracted features from the Arabic tweets and their corresponding counts.

3) Compiling the extracted features: The extraction of the features was followed by the coding step. To assist the automatic coding of a feature, we developed a LookupTable for each feature. The LookupTable is a binary table contains a unique occurrence of all possible values of that feature extracted from the Arabic ArSAS tweets. Each LookupTable is built by scanning its corresponding column(s) in the corpus and adding a unique occurrence value for all possible values of that feature.

To facilitate this final process, we developed a Graphical User Interface (GUI) to manage the compilation of each feature extracted from the tweets and assigning SAs to tweets. Table IV lists the different functions performed by the system and the numbers of the extracted features. The values of features are binary values located from the LookupTable. A value of 0 means that the feature does not exist in a tweet, otherwise it is 1. Fig. 2 depicts the GUI functions to be used to compile the extracted features into the final DataFile.csv.

TABLE. II. A SAMPLE OF THE FEATURES IN THE LOOKUPTAPLE EXTRACTED FROM THE ARABIC TWEETS

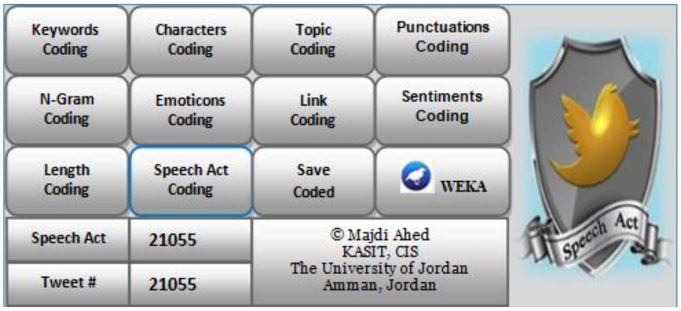
Keywords Features	Initial Words Features	Punctuations	Special Characters	N-Gram Features	Emoticons Features	Speech Acts	Topic	Sentiment Label
غانا	المباراة القادمة	"	#	كأس العالم	😫	Assertion	Event	Positive
شرم الشيخ	هل هذه	?	❕	شباب العالم	😰	Expression	Entity	Negative
تيران	وزير خارجية	!	✨	بسم الله	😌	Request	Long-Standing	Neutral
مصر	ومع السيسي		👎	افصل لاعب	😚	Question		Mixed
اوروبا	اهداف المباراة		⚽	حصار قطر		Recommendation		
الجزائر	طبعا 25		❤	ولي العهد		Miscellaneous		
محمد صلاح	منتدى		🏃	تفتح تحقيق				
قطر	يعني رايح		❓	المزيد من				
سويسرا	قولوا_ل_قطر_كبتين		✊	ثورة يناير				
مرتضى منصور	بسم الله		🏆	حصار اقتصادي				
الزمالك	هذا الربيع		👍	وزير خارجية				
الدنمارك	ملابس		👌	النجم المتالق	😊			
مصر	دول المال		👂	الدوري الانجليزي				
هدف	رحم الله		💖	الربيع العربي	😔			
السيسي	تصفيات كأس		💘	ببيعوا سمك	😟			
شكرا	ياه جه		🙌	العالم العربي	😎			

TABLE. III. SUMMARIZATION OF THE EXTRACTED FEATURES AND THEIR COUNTS EXTRACTED FROM THE ARABIC TWEETS

Feature Name	Count of Features
Punctuation	3
Twitter Special Chars	172
Topic	3
Sentiment	4
Emoticons	68
Keywords	1656
N-grams	2658
Speech Act categories	6
Link	1
Long	1

TABLE. IV. THE SYSTEM'S MAIN FUNCTIONS AND EXTRACTED FEATURES

Function	What it does	Number of features
Keywords Coding	Assigning values to keyword features	1656
Characters Coding	Assigning values to character features	172
Topic Coding	Assigning values to the topic features	3
Punctuation Coding	Assigning values to the punctuation features	3
N-Gram Coding	Assigning values to n-gram sequences	2658
Emoticons Coding	Assigning values to the existence of emoticons	68
Link Coding	Assigning values to the existence of hyperlinks	1
Sentiments Coding	Assigning values to the types of sentiments	4
Length Coding	Assigning a value to the length of a tweet	1
Speech Act Coding	Assigning a value to the SA type	6
Save Coded Data	Saving the compiled table of features as (DataFile.csv)	-
WEKA	Launching the Weka's package	-



Speech Act 21055

Tweet # 21055

Fig. 2. The GUI of the ArTSAC Corpus.

4) Generating the coded file: The final step in the process of developing the ArTSAC corpus was to generate the SA data file, which is a binary coded file containing all values of SA features. The final file is an Excel comma-separated file (".csv"), suitable to be processed by Weka's SVM algorithm. We called this file "DataFile.csv".

It is important to mention that the structure of DataFile.csv conforms to the structure of the dataset, which would be processed by Weka. This structure has a header of metadata, which is required by Weka to identify each attribute in the file. The header has labels such as *a*1, *a*2, *a*3,..., etc. where '*a*' stands for an attribute and the last column is labeled with '*c*', which contains the value of the SA class (c.f. Table I).

IV. DEVELOPMENT OF THE CLASSIFIER

A. Support Vector Machine

Support Vector Machine (SVM) lies under the category of supervised learning algorithms used for classification. SVM was originally designed to work when data has exactly two classes. In other words, it can be used with binary classification problems. The multiclass SVM problem aims to assign labels to instances, where the labels are drawn from a finite set of several elements.

The traditional approach to solving this problem using SVM is to reduce the single multiclass problem into several multiple binary classification problems. The most common technique in practice is to build one-versus-all classifiers and to choose the class which classifies the test instances with the greatest margin. Another strategy is to build a set of one-versus-one classifiers and to choose the class that is selected by the most classifiers. While this involves building classifiers, the time for training classifiers may actually decrease, since the training data set for each classifier is much smaller.

One way to solve the SVM training problem is to use sequential minimal optimization (SMO) [36, 37]. The setup parameters of SVM were gamma and kernel. Also, we used the C parameter to control the cost of misclassification on the training data. The best performance of SVM was when setting the kernel to “poly”, gamma to “auto”, and C to 1.

The annotated tweets and the extracted features that we obtained from the previous step were used to train the SVM classifier. The data was saved as a single Excel sheet named (DataFile.csv). In this research, we used Weka (Waikato Environment for Knowledge Analysis) machine learning software [35], which is developed at the University of Waikato, New Zealand. Weka’s SVM was implemented as a Java class that has properties. In this implementation, all missing values were replaced, and nominal attributes were transformed into binary ones. Furthermore, and by default, all attributes were normalized. This means that all output coefficients would be based on the normalized data rather than the original data. Such a step is very important for interpreting the results of the classifier.

For multiclass classification, we used Weka’s SVM which implemented a pairwise one-versus-one classification technique. The option that fits calibration models to the outputs of SVM is used to achieve accurate probability estimates. However, the predicted probabilities in the multi-class classification are coupled by using Hastie and Tibshirani's pairwise coupling method [39].

One advantage of using Weka is its flexibility of providing a set of alternatives to perform testing of the created classification model. These alternatives include use training set, supplied test set, cross-validation, and percentage split. In our experiments, we used the training set option to perform the testing, such that the training dataset (DataFile.csv) was also the test dataset. The output of training the classifier is a set of important measures which are: precision, recall, and F-score. Table V shows the results of running the SVM classifier on the ArTSAC dataset.

TABLE. V. THE PERFORMANCE EVALUATION OF THE SVM CLASSIFIER RUNNING ON THE ARTSAC DATASET

SA Category	Precision	Recall	F-Score
Assertion	0.963	0.855	0.894
Expression	0.882	0.966	0.922
Request	1.000	0.112	0.201
Question	1.000	0.065	0.123
Recommendation	0.911	0.809	0.857
Miscellaneous	1.000	0.083	0.154
Weighted Average	**0.906**	**0.903**	**0.896**

B. Evaluation of ArTSAC

Before we discuss the results we obtained from our modified ArTSAC corpus, we start with highlighting the previous results obtained from the ArSAS corpus [33] then we compare the results from running SVM on our modified ArTSAC corpus and compare it with the ArSAS corpus.

1) Features extractions in ArSAS and the modified ArTSAC: The features of ArSAS were extracted using the Farasa part-of-speech tagger [40], which has been modified to extract hashtags, emojis, and URLs. On the other hand, features such as unigrams, bigrams, and trigrams were extracted manually [34]. Our modified ArTSAC made benefits from all features in ArSAS in addition to the enhanced set of extracted features. Wherever applicable, the new features of ArTSAC were extracted automatically. Others were extracted manually. All ArTSAC features were extracted through a developed system as shown in Fig. 2.

2) The results of running SVM on the modified ArTSAC: Table V shows the results of running the SVM classifier on the modified ArTSAC corpus. Here we report the F-Score rate for each SA category. The Expression SA category achieved the highest F-Score with a rate of (92.2%). This was followed by the Assertion SA (89.4%), Recommendation SA (85.7%), Request SA (20.1%), Miscellaneous (15.4%), and Question SA (12.3%). However, the least number of tweets were in the Recommendation category (107 tweets) and the Miscellaneous category (60 tweets). The weighted average of all features achieved an F-Score rate of (89.6%).

3) Comparison between ArSAS and the modified ArTSAC: Here we report the comparison results of the SVM classifier running on the ArSAS dataset and the ArTSAC dataset. In the first experiment, we ran Weka’s SVM on ArTSAC using the same feature set of ArSAS [33], which include the following features:

- Lexical features: unigram, bigram, and trigram segments.
- Syntactic features: punctuation marks, twitter special characters, Emoticons, and hyperlinks.
- Structural features: tweet’s length, and part-of-speech (POS) tags.

4) In the second experiment, we ran Weka’s SVM using all features in ArTSAC. Table VI shows the comparison results of running SVM on both datasets: ArSAS and ArTSAC using the F-Score measure.

Table VI shows that the F-Score rate of running the SVM algorithm on ArTSAC using all compiled features is (89.6%), which outperformed the same algorithm running on the original ArSAS dataset with an F-Score rate of (86.2%). However, when we attempted to run SVM on our ArTSAC dataset using the same features as in the ArSAS dataset, we got an F-Score rate of (81.2%). The reason for getting a lower F-Score rate compared with the original ArSAS dataset (using the same features), could due to the following main reasons: (1) in our study we used Weka’s SVM algorithm, while in [34] we don't know exactly how they implemented their SVM algorithm, and (2) we used SMO to optimize the SVM training set along with tuning parameters (*kernel*, *gamma* and *C*), while in [34] it was not clear what parameters they used to tune their algorithm. Fig. 3 shows the F-Score results of running SVM on different Arabic speech acts datasets extracted from Arabic tweets.

The results in Fig. 3 were achieved by using all features in ArTSAC and picking the tweets that have 0.8 and above confidence scores. Then we directed the SVM classifier to use only four SA classes out of the six classes explained in Table I, which are: Assertion, Expression, Request, and Question. The reason for reducing the number of classes to four was because the number of tweets found under the other two classes was very few (c.f. Table I); Recommendation (107 tweets) and Miscellaneous (60 tweets). Therefore, we believe that they might negatively affect the performance of the SVM classification algorithm. We also believe that the better performance of the ArTSAC dataset was due to two main reasons: (1) The newly added features, mainly the sentiment label, indicated a great deal of association between a tweet sentiment label and its SA category, and (2) The careful manual annotation of the keywords as well as the extended *n*-gram segments (i.e. 4-gram, 5-gram and beyond) added more semantic concentration to the extracted features and took the tweets to a level beyond the bag of words.

TABLE. VI. RESULTS OF RUNNING SVM CLASSIFIER ON BOTH ARSAS AND ARTSAC DATASETS

Test	F-Score
SVM/original ArSAS [34]	0.862
SVM/ArTSAC using same features as in ArSAS [34]	0.812
SVM/ArTSAC using all compiled features in ArTSAC	0.896

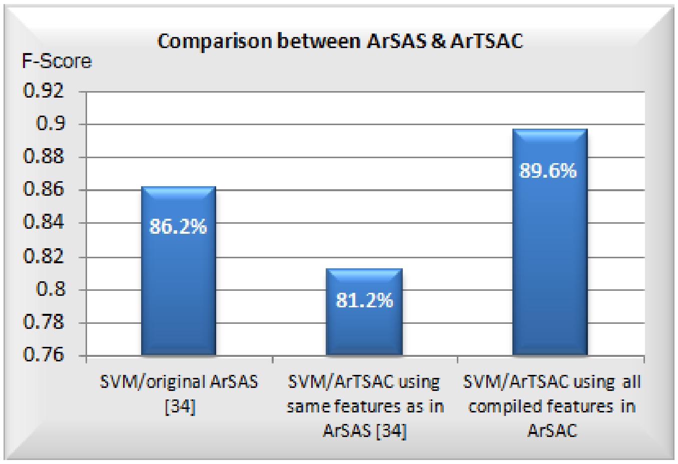


Fig. 3. F-Score Results of Running SVM on ArSAS and ArTSAC Datasets.

V. CONCLUSIONS

In this paper, we presented the development and construction of a richly annotated reference corpus of Arabic tweets for speech act classifications. The corpus, named ArTSAC, was built on top of a previous open-source modern standard Arabic twitter of SA corpus named ArSAS. ArTSAC inherited the features of ArSAS and added more annotated features before it has been put in practice with an SVM classification algorithm to classify Arabic tweets containing SAs.

The goal of ArTSAC is twofold: Firstly, to understand the purpose and intention of people's tweets which act in accordance with the SA theory, and hence positively influencing the development of many online applications. Secondly, as a future goal, to be used as a benchmark annotated corpus for testing and evaluating many Arabic software applications. ArTSAC has been put in practice to classify unseen Arabic tweets containing speech acts using the Support Vector Machine (SVM) classification algorithm. The results from our initial experimentation show that our developed corpus using the SVM algorithm achieved an average precision of (90.6%) and an F-score of (89.6%).

As for future work, we plan to use the ArTSAC corpus with deep learning based model for classifying speech-acts using a convolutional neural network (CNN).

ACKNOWLEDGMENT

The authors would like to thank the editor and the esteemed reviewers for their valuable comments to enhance the readability of the manuscript. Also they would like to thank the anonymous volunteers for doing the annotations and validation of the corpus.

REFERENCES

[1] A. S. Panah and M. M. Homayounpour, "Speech acts classification of Farsi texts," in 2008 International Symposium on Telecommunications, pp. 539-542. IEEE, 2008.

[2] X. Zhao, and J. Jiang, "An empirical comparison of topics in twitter and traditional media," Singapore Management University School of Information Systems Technical paper series. 2011.

[3] J. A. Austin, How to Do Things With Words, 2nd ed., Cambridge, Massachusetts, United States: Harvard University Press, 1975.

[4] X. Liao, Y. Huang, J. Wei, Z. Yu and G. Chen, "A Heterogeneous Graph Model for Social Opinion Detection," in International Conference on Machine Learning and Cybernetics ICMLC, Lanzhou, China, 2014.

[5] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," ACM Transactions on Asian Language Information Processing (TALIP), vol. 8, no. 4, pp. 1-22, 2009.

[6] G. Xu, H. Lee, M. W. Koo, and J. Seo, "Convolutional Neural Network using a threshold predictor for multi-label speech act classification," in 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), Jeju, South Korea, pp. 126-130. IEEE, 2017.

[7] D. Kim, H. Kim, and J. Seo, "Speech Act Classification Based on Individual Statistical Models in a Multi-Domain," in The 16th IEEE International Symposium on Robot and Human Interactive Communication, Jeju, South Korea, pp. 845-847. IEEE, 2007.

[8] H. Xuefeng. and Z. He, "Methods and characters of speech acts in online shopping," in 2012 IEEE Symposium on Robotics and Applications (ISRA), Kuala Lumpur, pp. 416-418. IEEE, 2012.

[9] F. Al-Hindawi and H. Al-Masudi, "The Speech Act Theory in English and Arabic," Open Journal of Modern Linguistics, vol. 4, pp. 27-37, 2014.

[10] Y. Ren and J. Tian, "Sentiment Analysis of Internet Performance Data," in 2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, pp. 622-628. IEEE, 2017.

[11] S. Zamani, M. Asadpour and D. Moazzami, "Rumor Detection for Persian Tweets," in 2017 Iranian Conference on Electrical Engineering (ICEE), Tehran, Iran, pp. 1532-1536. IEEE, 2017.

[12] W. Alabbas, H. M. Al-Khateeb, and A. Mansour, "Arabic Text Classification Methods: Systematic Literature Review of Primary

Studies," in 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), Tangier, Morocco, pp. 361-367. IEEE, 2016.

[13] N. Abdelhade, T. Hassan, A. Soliman, and H. Ibrahim, "Detecting Twitter Users' Opinions of Arabic Comments During Various Time Episodes via Deep Neural Network," in International Conference on Advanced Intelligent Systems and Informatics, Cairo, Egypt, pp. 232-246. Springer, Cham, 2017.

[14] M. Jeong, C.Y. Lin, and G. G. Lee, "Semi-Supervised Speech Act Recognition in Emails and Forums," in 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, vol. 3, pp. 1250-1259, 2009.

[15] V. Carvalho and W. Cohen, "Improving "email speech acts" Analysis via n-gram selection," in HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech, NY, USA, pp. 35-41, 2006.

[16] W. Cohen, V. Carvalho and T. Mitchell, "Learning to Classify Email into Speech Acts," in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, pp. 309-316, 2004.

[17] B. Bayat, C. Krauss, A. Merceron and S. Arbanowski, "Supervised Speech Act Classification of Messages in German Online Discussions," in Twenty-Ninth International Florida Artificial Intelligence Research Society Conference, Florida, USA, 2016.

[18] C. Moldovan and V. Rus, and A. C. Graesser, "Automated Speech Act Classification for Online Chat," in The 22nd Midwest Artificial Intelligence and Cognitive Science Conference, Cincinnati, USA, pp. 23-29, 2011.

[19] V. Rus, C. Moldovan, N. Niraula and A. C. Graesser, "Automated Discovery of Speech Act Categories in Educational Games," in The 5th International Conference on Educational Data Mining Society, Chania, 2012.

[20] R. Zhang, D. Gao and W. Li, "What Are Tweeters Doing: Recognizing Speech Acts in Twitter," in Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence, San Francisco, USA, 2011.

[21] R. Zhang, D. Gao and W. Li, "Towards scalable speech act recognition in Twitter: tackling insufficient training data," in Proceedings of the Workshop on Semantic Analysis in Social Media, Avignon, pp. 18-27, 2012.

[22] S. Vosughi and D. Roy, "Tweet Acts: A Speech Act Classifier for Twitter," in The Tenth International AAAI Conference on Web and Social Media, Cologne, Germany, 2016.

[23] D. Nemer, "Celebrities Acting up: A Speech Act Analysis in Tweets of Famous People," Journal of Social Networking, vol. 5, no. 1, pp. 1-10, 2016.

[24] S. C. Herring, A. Das, and S. Penumarthy, "CMC Act Taxonomy," 2005. [Online]. Available: http://info.ils.indiana.edu/~herring/cmc.acts, [Accessed Feb. 17, 2020].

[25] M. Ashcroft, A. Fisher, L. Kaati, E. Omer and N. Prucha, "Detecting jihadist messages on twitter," in European Intelligence and Security Informatics Conference, Manchester, UK, pp. 161-164. IEEE, 2015.

[26] A. A. Elmadany, S. M. Abdou and M. Gheith, "Recent Approaches to Arabic Dialogue Acts Classifications," in The 4th conference of Natural Language Processing, Sydney, Australia, vol. 5, no. 4, pp. 117-129, 2015.

[27] L. Shala, V. Rus, and A. C. Graesser, "Automated Speech Act Classification in Arabic," Subjectivity and Cognitive Processes, vol. 14, no. 2, pp. 284-292, 2010.

[28] M. Hijjawi, Z. Bandar and K. Crockett, "User's Utterance Classification using Machine Learning for Arabic Conversational Agents," in 5th International Conference on Computer Science and Information Technology, Amman, Jordan, pp. 223-232. IEEE, 2013.

[29] M. Hijjawi, Z. Bandar, K. Crockett, and D. Mclean, "ArabChat: An Arabic Conversational Agent," in 6th International Conference on Computer Science and Information Technology (CSIT), Amman, Jordan, pp. 227-237. IEEE, 2014.

[30] ElA'awar, "Verbal Actions in Surat Al-Kahf - A Deliberative Study," Department of Linguistics, Faculty of Arts and Sciences, University of Mentor, Algeria, Algeria, 2011.

[31] M. Medawar, "Verbal Actions in the Holy Quran (Surat Al-Baqara)- A Deliberative Study," Department of Arabic Language, College of Arts and Sciences, Hajj Lakhdar University, Algeria, Algeria, 2014.

[32] F. A. M. Jawad, "A Pragmatic Analysis of Illocutionary Speech Acts in Standard Arabic with a Special Reference to Al-Ashter s 'Epistle'," Journal of University of Babylon 19, no. 4, pp. 606-625, 2011.

[33] A. Elmadany, H. Mubarak and W. Magdy, "ArSAS: An Arabic Speech-Act and Sentiment Corpus of Tweets," in 3rd Workshop on Open-Source Arabic Corpora and Processing Tools, Miyazaki, Japan, p. 20, 2018.

[34] B. Algotiml, A. Elmadany, and W. Magdy, "Arabic Tweet-Act: Speech Act Recognition for Arabic Asynchronous Conversations," in In Proceedings of the Fourth Arabic Natural Language Processing Workshop, pp. 183-191. 2019.

[35] G. Holmes, A. Donkin and I. H. Witten, "Weka: A machine learning workbench," in Proceedings of ANZIIS'94-Australian New Zealand Intelligent Information Systems Conference, pp. 357-361. IEEE, 1994.

[36] Nabble, "Explanation of SMO Parameters?," [Online]. Available: http://weka.8497.n7.nabble.com/Explanation-of-SMO-Parameters-td21768.html, [Accessed Feb. 17, 2020].

[37] J. Platt, "Fast training of support vector machines using sequential minimal optimization," Advances in Kernel Methods-Support Vector Learning, AJ, MIT Press, Cambridge, MA, pp. 185-208, 1999.

[38] J. R. Searle, Expression and Meaning: Studies in the theory of speech acts, Cambridge University Press, 1985.

[39] T. Hastie, and R. Tibshirani, "Classification by pairwise coupling," in Advances in neural information processing systems, pp. 507-513. 1998.

[40] K. Darwish, and H. Mubarak, "Farasa: A new fast and accurate Arabic word segmenter," in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 1070-1074. 2016.

Recognition of Image in Different Cameras using an Improved Algorithm in Viola-Jones

Washington Garcia - Quilachamin[1]
Student PhD, UNMSM, Lima – Peru
Professor Faculty of Engineer
ULEAM, Manta - Ecuador

Luzmila Pro - Concepción[2]
Professor Faculty of Engineer
UNMSM
Lima – Peru

***Abstract*—Technological evolution through computer tools has given rise to tasks of impossible recognition for an ordinary man, but at the same time favorable for the safety of people. Deep learning is considered a tool that uses images and video to detect and interpret real-world scenes. Therefore, it is necessary to validate the application of an algorithm with different cameras for the recognition of people, being a contribution to surveillance in domestic environments and of companies. In this research, an algorithm is presented that, through a camera, allows to detect the image of a person. The objective of this research is to validate the process in the recognition of the image with four cameras through the application of the improved algorithm Viola-Jones. The validation was carried out through a mathematical analysis, which allowed us to base the recognition of the image using four different cameras. As a result of the study, an effective and functional validation was obtained, about the results achieved with the application of the algorithm, using the four cameras and effective in the speed-based recognition concerning the different tests performed on the capture and recognition of each image, reducing the recognition time and optimizing the software and hardware used.**

Keywords—Algorithm; video surveillance; cameras; image of a person

I. Introduction

With technological evolution, the application of algorithms with cameras for the recognition of objects, people or events are part of the studies of artificial vision and deep learning, which is of much interest research, in the innovation of real-world scenes. Therefore, recognition tasks are considered complex for a computer, according to [1], [2]. The authors [3] consider that from the study of the recognition of objects, innovation in facial and people recognition is started.

The authors in [4] proposes that every process, algorithm, model, etc., is based on the effective recognition of human activity (HAR) which consists of understanding what people are doing from their position, figure, movement or other space-time information derived from their video sequences. according to [5], [6] and [7], Some factors incentive this research with the recognition of a person's image is based on the recognition of different types of images that relate to human movements such as running, limping, jumping, etc.

So the authors in [8], establish that the Viola-Jones algorithm is currently one of the most used to solve search problems considering the face of a person. According to [9], [10] pattern recognition systems are developed as a part of continuous monitoring of human behavior in the area of assisted living, rehabilitation, and entertainment.

The authors in [11], [12] and [13] mention that applications may vary in object recognition systems in a video surveillance system located at an intersection of roads, restaurants, education centers, shopping centers or any other public meeting place.

For what is considered the field of security as a topic of broad action in offices, homes, and smart cities. Some authors [14], express that the security of a people and family is very important by considering that human pattern recognition systems can be connected to through device automatic.

The purpose of this research is to validate the process in the recognition of image applying the improved algorithm. The analysis was carried out through mathematical equations considering a definite integral from a zero point, which allowed us to justify the research carried out about the recognition of image of a person through four cameras, considering as parameters the time and distance from the camera, to the person's position inside the area.

As a result of the study in using an improved algorithm in viola - jones, for the recognition of image, using four cameras, data were obtained relating to the time and distance in which the image is detected with each camera. These data were used in the analysis, through the application of equations and definite integral, considering as reference the author [15]. With the effective use of the algorithm in a distance of 3m to 9m, the image capture and recognition as true positive, false positive, false negative is determined, considering that, at a greater distance of an image concerning the camera, the angle is smaller in recognition of image, otherwise its result related to recognition would be false negative.

This paper is structured as follows:

Section II shows the related work, which explains a previous review of the main models of image recognition over time, details the methodology used, the framework used to construct the proposed algorithm, improved algorithm, criteria, and the mathematical analysis applied.

Section III the results of the application of improved the algorithm are shown, and the analysis in the recognition of image of a person with the four cameras. Finally, the conclusions reached in the present study are detailed, and the future work.

II. Method and Related Works

The authors in [16], [17], [18], manifest that there are several studies on base at the recognition of an image using algorithms.

Convolutional Neural Network (dCNN) [19], Hidden Markov (HMM) [20], [21],[22] and Support Vector Machines (SVM) [23], as the authors argue, are algorithms models used to classify images, the analyze data, with computational methods for the patterns recognition.

The models described are based on algorithms that can identify images differently, as they consider it [7], [24], [25], , who mention studies related to the recognition of faces, people, traffic signs, tumors and many other aspects of visual data in different activities that the human performs.

Of all the models, the SVM was chosen to be fully implemented in Matlab, because in real-time it detects people not included in an upright position. Also, this model is related to classification and regression problems; this will be used in our research through the algorithms described.

The SVM model is used to perform recognition of people in a real application [23], and with the criteria of the authors of [8], the implementation of the Viola-Jones algorithm allowed us to obtain the capture of the image of a person. Which is represented with a frame, and that these pixels minimizing the amount and time of necessary calculations.

In [8] and [26], the authors state that the Viola-Jones algorithm is based on a series of classifiers and employs a method of approximation based on appearance. It is divided into two stages: a first stage of learning the classifier based on a large number of positive examples and negative examples, and a recognition phase by applying this classifier in relation to the images not known.

These features were used in our research and that allowed us to apply efficiently from an integral image in time real considering a person for this study.

A. *Methodology*

To analyze the algorithm's execution based on the recognition of people, our research applies an exploratory methodology, considering the data collection of four video surveillance cameras, where 840 was obtained captures, corresponding to 210 images per camera. This study is carried out to determine the effectiveness in the execution of the algorithm viola jones based on the recognition of image.

B. *Framework of Work*

The purpose of this research is to understand and validate the process to recognize an image based on the using an improved algorithm in Viola-jones, using the image acquisition tool, computer vision system, and deep learning. the procedure is described through the framework shown in Fig. 1.

The improved algorithm will be used to capture a pattern referring to the image of a person, considering that the objective is to validate the recognition of image in real-time through the four cameras hilook, toshiba, max, and turbo. the calculation of the data about the integral image is expressed in [8].

With the information obtained concerning parameters time and distance, in the recognition of the image, the respective calculations and analysis were developed. These are fundamental in the validation of the algorithm, considering the authors [15] and [27].

C. *Algorithm Applied*

The algorithm described below was improved, based on Viola-Jones and has been used in the recognition of image in real-time.

```
Data: image, recobj, imqrec, imshow, frame, bboxes, r
image=vision. recobj ( );
 recog=imaq.VideoDevice ( );
set (recog);
   while (true)
       frame=step(recog);
       bboxes=step (facerecognition,frame);
       imqrec=insertObjectAnnotation(frame, ,bboxes,);
       imshow(imqrec)
       exit of imshow
       r=findobj();
    end
end
```

The data obtained that refers to the recognition are imported and saved in the routing path.

D. *Viola-Jones*

The purpose of this research is to determine an algorithm based on Viola-Jones that allows us to recognize an image and validate your process through the mathematical equation (1). Whereas this work is motivated in the recognition of faces on real-time video, according to [28].

$$I(x,y) = \sum_{xi<x,yi<y} i(x',y') \tag{1}$$

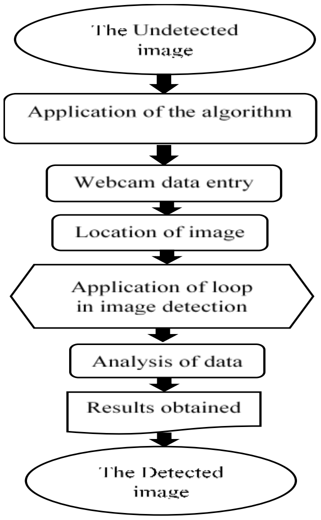


Fig. 1. Framework for Image Recognition.

Where: (x, y) is the integral image calculated in pixels, (x', y') is the original image. Using the integral image, any sum of a rectangular area ABCD can be calculated efficiently, (2):

$$\sum_{(x,y)\in ABCD} i(x', y') = \mathrm{I(D)} + \mathrm{I(A)} - \mathrm{I(B)} - \mathrm{I(C)} \qquad (2)$$

E. Tools used

For this study, the improved algorithm in the recognition of image is used and four different types of cameras. These were selected by their size and class.

Among the cameras selected and which were used for the proposed objective, is the computer's webcam and cameras for infrared security type Dome and Bullet. The software that was used for testing and algorithm improvement was Matlab 2017b.

According to [29], the authors describe the parameters of cameras considered in this study and shown in Table I.

640 x 480 pixels were considered for the image resolution of the Toshiba camera because it does not support its configuration, while the image resolution on the Hilook, Max and Turbo cameras is assigned by default when used with the Easy Cap device, which allows the conversion of the image.

F. Mathematical Procedure

The mathematical procedures are based, considering [29], and for the development of this process in our research, it is considered the operation of vectors, matrices, and definite integrals, the equation of the slope (1), was applied the equation mathematical concerning the function f (x) (2) and its derivative as a function of time (3).

$$\mathrm{m}=\frac{y_2-y_1}{x_2-x_1} \qquad (2)$$

$$\mathrm{f(x)}=\frac{\mathrm{y_2-y_1}}{\mathrm{x_2-x_1}} = \frac{\mathrm{dy}}{\mathrm{dx}}\ \mathrm{dt} \qquad (3)$$

$$\mathrm{f(x)}=\frac{\mathrm{dy}}{\mathrm{dx}}\mathrm{dt} \qquad (4)$$

Where “dy”, represents recognition angle, “dx” distance and “dt”, concerning the time, which is considered as (x, y) respectively.

The equation of definite integral (4) was considered, which was applied to evaluate the values obtained concerning the time and distance parameters of each camera used for the recognition. Starting from a zero point.

$$\int_0^d \frac{1}{2}(x/t).dx \qquad (4)$$

where “d” will be the distance value and “t” is the time value obtained from each camera, respectively.

TABLE. I. CHARACTERISTICS OF CAMERAS

Camera	Night Vision	Image format	Image resolution	Iris opening
Toshiba	Negative	YUY2	640 x 480	Permanent
Hilook	Positive	YUY2	720 x 576	Automatic
Max	Positive	YUY2	720 x 576	Automatic
Turbo	Positive	YUY2	720 x 576	Automatic

III. RESULTS AND ANALYSIS

A. Results of the Recognition by Cameras

In Table II, the person's height, recognition angle, camera height, and distance are described, which are the parameters used and the data obtained concerning the recognition of an image, in this case, concerning a person's.

Table III shows the data obtained by the four cameras, considering the time in seconds when recognition the image of a person concerning distance.

The data of Table III were applied in Fig. 2, which are the result of the execution of the algorithm when using the four different cameras in recognition of the image of a person. Lines shown in a different layout, are related to time and distance, concerning each camera used in this research when detecting the image.

TABLE. II. MEASUREMENT PARAMETERS

Person stature (m)	Recognition angle	Camera height (m)	Distance (m)
1.78	16°	0.88	3.00
1.78	14°	0.88	4.00
1.78	12°	0.88	5.00
1.78	10°	0.88	6.00
1.78	8°	0.88	7.00
1.78	6°	0.88	8.00
1.78	4°	0.88	9.00

TABLE. III. RECOGNITION TIME BY CAMERA

Distance (m)	Types of cameras			
	Hilook (s)	Max (s)	Toshiba (s)	Turbo (s)
3.00	0,5008767	0,4952986	0,5842529	0,4894874
4.00	0,4786243	0,4950434	0,6029704	0,5046813
5.00	0,4836451	0,4976678	0,5764817	0,4875867
6.00	0,5693954	0,5144184	0,5583289	0,4715800
7.00	0,4846173	0,4977600	0,6227089	0,5037189
8.00	0,4968127	0,4843582	0,5801745	0,5045335
9.00	0,4921652	0,4825438	0,5648531	0,4648238

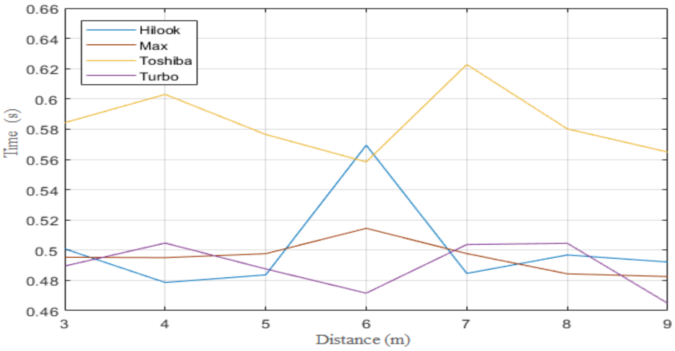


Fig. 2. Time-Distance Variations.

Fig. 2, shows the line concerning the Hilook camera, which presents variations, and the showing the largest peak of all the lines at 6m and low speeds in the intervals of 4m and 5m., for what is considered that this camera is less efficient in the distance range of 3m to 9m. It also shows the stroke that represents the Max camera, this stroke represents the most continuous speeds of all data and It has a slight drop peak in the 6m interval and represents the second-fastest.

The strokes, of Toshiba Webcam camera, are progressively scaling, presenting a slight peak in 4m., And the second peak of fall, this time lighter in 7m. Even so, the speed values in the last intervals in this camera are slower, but it maintains its continuity. The stroke concerning the Turbo camera is the fastest in most intervals except in 4m, which shows a slight peak of lower speed, also in the interval of 7m and 8m, the same problem occurs. It is slightly slower. Therefore, this camera is the fastest compared to the strokes of the previous cameras.

The results obtained allowed to demonstrate the performance in the speed of the different cameras in the recognition of image, considering that the proposal is not to evaluate the cameras, but the effectiveness using the improved algorithm in viola-jones. Considering this result, the algorithm works, regardless of the type and brand of the cameras.

Table IV shows the image recognition framed in a frame (yellow contour), in this case, concerning a person's, it is considered the distance of a person regarding the camera. The images were captured using four different cameras through the application of the algorithm at different distances and positions in real-time, of course, the values vary according to their specifications.

The application of the algorithm in the recognition of image it allows us to validate its effectiveness in the four cameras. Table V shows the results of the recognition at different focal angles.

B. Analysis in the Recognition of the Image

In this case, concerning a person's. The results obtained in the recognition of the image in the four cameras, considering a distance of 3, 4, 5, 6, 7, 8 to 9m, and with the application of the algorithm configured in Matlab. It was determined that the Turbo camera detects positively the image in the established range.

The Toshiba Webcam camera, detects the image positively in the range of distance of 3m to 6m and false positive at a distance of 7m, positively detecting again in the range of 8m to 9m.

With the Max camera, the recognition of image is false negative at a distance of 3m, considering that in the range of 4m to 9m the recognition is positive.

Finally, with the Hilook camera, image recognition is positive at a distance of 3m, 5m and in a range of 7m to 9m. In this recognition there is a false negative at a distance of 4m and a false positive at 6m. The statistical table of Fig. 2 is also considered, concerning the distance with the time in image recognition, which allows establishing the reliable Turbo HD camera in the application.

C. Results with the Mathematical Equation

The speed is the result in the recognition of the image, concerning distance and time. Table VI shows the results obtained, through the developed calculations considering 0m as a lower limit, up to an upper limit of 3m up to 9m.

TABLE. IV. IMAGE RECOGNITION

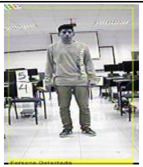
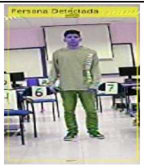
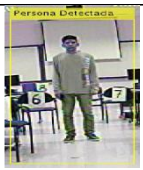
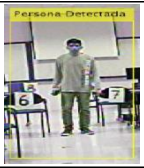
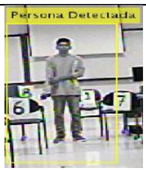
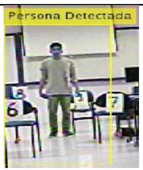
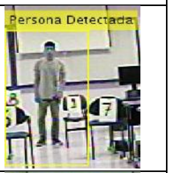
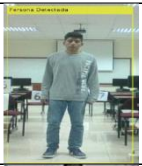
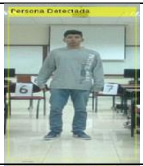
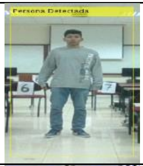
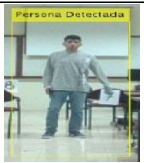
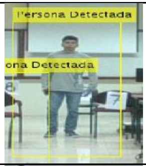
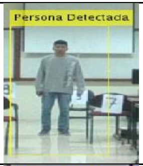
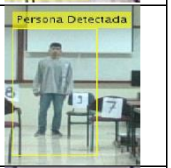
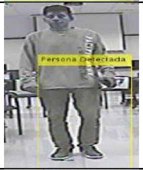
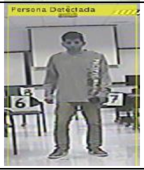
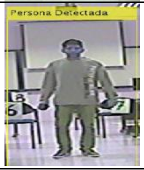
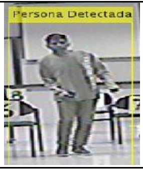
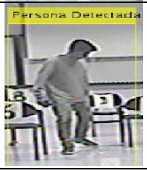
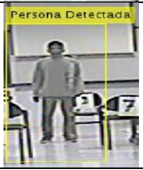
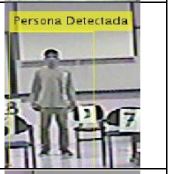
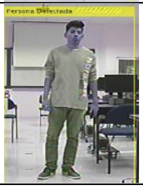
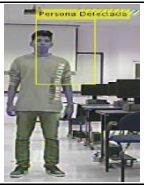
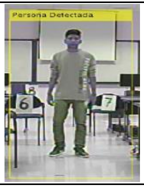
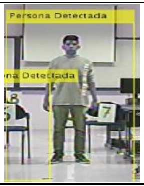
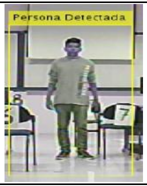
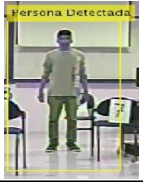
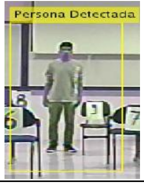
Camera type	Distance						
	3m	4m	5m	6m	7m	8m	9m
Turbo							
Toshiba							
Max							
Hilook							

TABLE. V. FOCAL ANGLE AND IMAGE RECOGNITION

Camera	**Focal angle**		**Recognition**
	Inclination	**Rotation**	
Turbo	0º to 90°	0º to 360°	Positive
Toshiba	0º to 180º	0º	Positive / False-Positive
Max	0º to 180º	0º to 360º	False-Negative / Positive
Hilook	0° to 180°	0º to 360º	Positive / False-Negative False-Positive

TABLE. VI. CAPTURE SPEED

Distance (m)	**Types of cameras**			
	Hilook (m/s)	**Max (m/s)**	**Toshiba (m/s)**	**Turbo (m/s)**
3.00	8,98424702	9,08542847	7,70214405	9,19329078
4.00	16,71457132	16,16019929	13,26764962	15,85158792
5.00	25,84539779	25,11715646	21,68325551	25,63646629
6.00	31,61247878	34,99097233	32,23906196	38,16955766
7.00	50,55535574	49,22050788	39,34422649	48,63823851
8.00	64,41059176	66,06680758	55,15581950	63,42492619
9.00	82,28944265	83,93020489	71,70005794	87,12978983

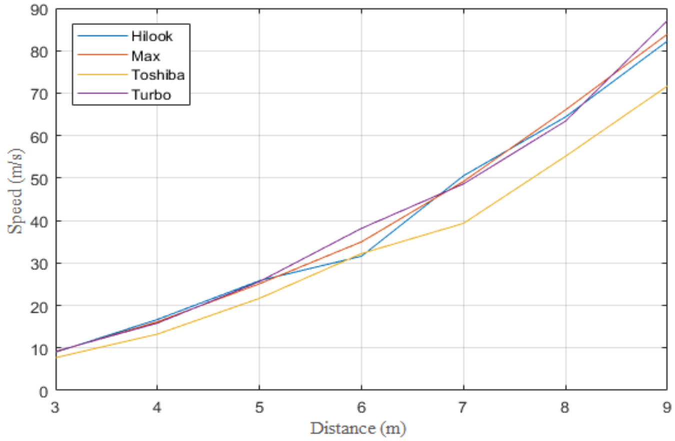


Fig. 3. Speed – Distance.

Fig. 3 shows the strokes obtained between the ratio of the speed and the distance, of the four different cameras used for the recognition of the image of a person.

D. Analysis with the Applied Mathematical Equation

Based on the results obtained in Table VI, the speed of the recognition of the image of a person, in a distance range of 3m to 9m. it was determined: that the average speed in recognition by the Hilook camera is 40.058m/s. The Max camera detects the image of the person at an average speed of 40.653m/s. With the Toshiba camera, recognition average speed was of 34.441m/s, and finally, with the Turbo camera, the recognition of the image in a range of 3m to 9m was with an average speed of 41.149m/s.

Fig. 3, shows the data the Table VI, in this figure shows the lines concerning the speed and distance parameters of the four cameras. Its lines present variations to the speed, in relation to the recognition of the image in a range of 3m to 9m. Of course, the values vary according to their specifications and this determines the Toshiba camera to be the most optimal because of its average speed of 34.441m/s, but it must be considered that this camera works with the computer processor.

IV. CONCLUSION

The use of the improved algorithm in viola-jones, allowed to obtain an effective and functional evaluation concerning 840 captures, corresponding to 210 images per camera, achieved in the recognition of the image of a person in the four cameras. With the data of time, distance and speed applied through the mathematical equation, it allowed determining that one of the four cameras used in this research to show a higher speed than the others, demonstrating that this is the fastest at the time of capturing and recognition of the image of a person, but this camera works with the computer processor.

So that, by applying the algorithm in the cameras, it was possible to validate its efficiency and speed in the recognition of image, considering that this recognition is reduced concerning time, optimizing the software and hardware used.

In this research, the hilook camera is considered the most suitable with an optimal displacement based on the speed of 40.058m / s. This also allows validating the effective in the recognition of the image of a person in the distance range of 3m to 9m, and the results obtained and analyzed strengthen the research developed concerning the using improved algorithm in viola-jones applied in the recognition of the image of a person.

Its validation and subsequent implementation of video surveillance cameras with the improved algorithm will serve as a contribution to the security and surveillance of people in domestic and business areas.

V. FUTURE WORK

This study is considered in the future to link concerning improving energy efficiency, video surveillance and reducing the CO2 that is affecting climate change.

REFERENCES

[1] X. J. Lin, Q. X. Wu, X. Wang, Z. Q. Zhuo, and G. R. Zhang, "People recognition in multi-cameras using the visual color processing mechanism," Neurocomputing, vol. 188, pp. 71–81, 2016.

[2] Mathworks, "Image Category Classification Using Bag of Features," 2016.

[3] K. Sinhal, "Object Detection using Deep Learning for advanced users," 2017.

[4] S. Sun, J. Zhao, and Q. Gao, "Modeling and recognizing human trajectories with beta process hidden Markov models," Pattern Recognit., vol. 48, no. 8, pp. 2407–2417, 2015.

[5] C. Benabdelkader, R. Cutler, H. Nanda, and L. Davis, "EigenGait : Motion-Based Recognition of People Using Image Self-Similarity," pp. 284–294, 2001.

[6] J. Matai, A. Irturk, and R. Kastner, "Design and implementation of an FPGA-based real-time face recognition system," Proc. - IEEE Int. Symp. Field-Programmable Cust. Comput. Mach. FCCM 2011, pp. 97–100, 2011.

[7] H. Xu, L. Li, M. Fang, and F. Zhang, "Movement human actions recognition based on machine learning," Int. J. Online Eng., vol. 14, no. 4, pp. 193–210, 2018.

[8] M. V. Alyushin and A. A. Lyubshov, “The Viola-Jones algorithm performance enhancement for a person’s face recognition task in the long-wave infrared radiation range,” Proc. 2018 IEEE Conf. Russ. Young Res. Electr. Electron. Eng. ElConRus 2018, vol. 2018-Janua, pp. 1813–1816, 2018.

[9] M. Y. Santos et al., “A Big Data Analytics Architecture for Industry 4.0,” WorldCIST 2017 Recent Adv. Inf. Syst. Technol., vol. 4, 2017.

[10] H. F. Nweke, Y. W. Teh, M. A. Al-garadi, and U. R. Alo, “Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges,” Expert Syst. Appl., vol. 105, pp. 233–261, 2018.

[11] X. Jiang, J. Sun, H. Ding, and C. Li, “A silhouette based novel algorithm for object detection and tracking using information fusion of video frames,” Cluster Comput., vol. 0, 2018.

[12] F. H. K. Zaman, M. H. Ali, A. A. Shafie, and Z. I. Rizman, “Efficient human motion detection with adaptive background for vision-based security system,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 7, no. 3, pp. 1026–1031, 2017.

[13] A. E. Papadakis, E. Tsalera, and M. Samarakou, “Survey on sound and video analysis methods for monitoring face-to-face module delivery,” Int. J. Emerg. Technol. Learn., vol. 14, no. 8, pp. 229–240, 2019.

[14] Ilhan Aydin and Nashwan Adnan Othman, “A New IoT Combined Face Detection of People by Using Computer Vision for Security Application,” Int. Artif. Intell. Data Process. Symp. (IDAP), Mal., pp. 0–6, 2017.

[15] G. D. Dean, Advanced engineering mathematics with matlab, 4th Editio. 2016.

[16] F. Kamaruzaman and A. A. Shafie, “Recognizing faces with normalized local Gabor features and Spiking Neuron Patterns,” Pattern Recognit., vol. 53, pp. 102–115, 2016.

[17] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu, “Silhouette analysis-based gait recognition for human identification,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 12, pp. 1505–1518, 2003.

[18] X. J. Wang, “A human body gait recognition system based on fourier transform and quartile difference extraction,” Int. J. Online Eng., vol. 13, no. 7, pp. 129–139, 2017.

[19] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, “Multi-type attributes driven multi-camera person re-identification,” Pattern Recognit., vol. 75, pp. 77–89, 2018.

[20] A. V. Nefian and M. H. Hayes, “Face detection and recognition using hidden Markov models,” pp. 141–145, 2002.

[21] A. Mahapatra, T. K. Mishra, P. K. Sa, and B. Majhi, “Human recognition system for outdoor videos using Hidden Markov model,” AEU - Int. J. Electron. Commun., vol. 68, no. 3, pp. 227–236, 2014.

[22] L. Speech, Processing, Daniel, H. James, and Martin, “Hidden Markov Models ch9,” no. Chapter 20, 2017.

[23] Y. Lu, K. Boukharouba, J. Boonært, A. Fleury, and S. Lecœuche, “Application of an incremental SVM algorithm for on-line human recognition from video surveillance using texture and color features,” Neurocomputing, vol. 126, pp. 132–140, 2014.

[24] A. Baca, Methods for Recognition and Classification of Human Motion Patterns – A Prerequisite for Intelligent Devices Assisting in Sports Activities, vol. 45, no. 2. IFAC, 2012.

[25] M. Kleinsmith, “Zero to Hero: Guide to Object Detection using Deep Learning: Faster R-CNN,YOLO,SSD,” 2016. [Online]. Available: http://cv-tricks.com/object-detection/faster-r-cnn-yolo-ssd/.

[26] P. Viola and M. Jones, “Robust real-time face detection,” Proc. Eighth IEEE Int. Conf. Comput. Vision. ICCV 2001, vol. 2, pp. 747–747, 2001.

[27] L. Yolanda, M. Martín, L. Gutiérrez Mendoza, L. Mary, and A. Nieves, “Guidelines to design of virtual learning objects (VLO). Application to the Teaching-Learning Process of Area under the Integral Calculus Curve Lignes directrices pour la conception d’ objets d’apprentissage virtuels (OAV). Application au processus d’ ense,” Rev. Científica Gen. José María Córdova, vol. 14, pp. 127–147, 2016.

[28] R. Klette, Concise Computer Vision. 2014.

[29] Garcia-Quilachamin W. Concepción L.P. Herrera-Tapia J. Salazar R.J.Toala-Mero W., “Validation of an algorithm for the detection of the image of a person using multiple cameras,” in Applied Technologies. ICAT 2019. Communications in Computer and Information Science, vol 1194. Springer, Cham, 2020, pp. 486–501.

An Improved RDWT-based Image Steganography Scheme with QR Decomposition and Double Entropy

Ke-Huey Ng[1], Siau-Chuin Liew[2], Ferda Ernawan[3]
Faculty of Computing, Universiti Malaysia Pahang
Gambang, Kuantan, Pahang Darul Makmur, Malaysia

***Abstract*—This paper introduces an improved RDWT-based image steganography with QR decomposition and double entropy system. It demonstrates image steganography method that hides grayscale secret image into grayscale cover image using RDWT, QR decomposition and entropy calculation. The proposed scheme made use of the human visual system (HVS) in the embedding process. Both cover and secret image are being segmented into non-overlapping blocks with identical block size. Then, entropy values generated from every image block will be sorted from the lowest value to the highest value. The embedding process starts by embedding the secret image block with lowest entropy value into the cover image block with lowest entropy value. The process goes on until all image blocks have been embedded. Embedding secret image into cover image according to the entropy values causes differences that HVS can less likely to detect because of the small changes on image texture. By applying the double entropy system, proposed scheme managed to achieve a higher PSNR value of 60.3773 while previous work gave a value of 55.5771. In terms of SSIM value, proposed scheme generated a value of 0.9998 comparing to previous work's value of 0.9967. The proposed scheme eliminated the false-positive issue and required low computational time of only 0.72 seconds for embedding and 1.14 seconds for extraction process. Also, it has shown better result compared to previous work in terms of imperceptibility.**

***Keywords*—Steganography; image steganography; transform domain; Redundant Discrete Wavelet Transform (RDWT); QR decomposition; entropy; human visual system (HVS); imperceptibility**

I. Introduction

Steganography [1] becomes more and more important as many people joined the cyberspace revolution that involves information exchanging technology. It is the science of information hiding. Its purpose is to convey a message without letting the existence of message being discovered except for the intended receiver, and if being discovered, the message is hard to be detected and recovered. Digital pictures, audio and video are increasingly furnished with distinguishing but imperceptible marks [2], which may contain a hiding copyright serial number or notice. This may directly help to prevent the unauthorized use.

In the context of image steganography, there are two (2) domains which are spatial and transform domain. Spatial domain involves the direct bitwise manipulation whereas transform domain focuses on the transformed image manipulation, which means the original cover image will be changed or transformed first before embedding secret message. Spatial domain is easier to be developed as compared to transform domain. It requires shorter computational time. However, it is more vulnerable to attacks. The reason is that it embeds secret information into cover image directly without transforming the cover image itself, this can cause the secret information to be destroyed easily if the stego image has been attacked.

Thus, transform domain is more preferable because it ensures a certain level of robustness as it withstands against attacks such as geometric attacks and compression. The secret information should still be present and can be detected regardless of the attacks done to stego image. Among different types of transform domain techniques, wavelet transform requires less computational cost compared to DCT and FFT (Fourier Transform) and offers sub-representations of the image that can be considered related to how the human visual system (HVS) perceives images. Generally, the wavelet transform allows embedding data in high frequency regions where the HVS cannot distinguish modifications compared to uniform regions with low frequency.

On the other hand, to avoid unauthorized users attacking stego image easily, the concept of entropy will be applied. It allows the embedding process to be done randomly on image blocks based on the calculation of entropy instead of sequentially placing the secret information from certain pre-set location to another location. This approach enhances the imperceptibility of secret information by embedding the information in image blocks with lower entropy values as they appeared to be less sensitive for HVS. Also, by using this approach, the risk of embedded information being fully attacked or damaged would be lowered because it spreads the secret information randomly on the cover image.

Also, the existing algorithms only do the embedding and extraction process without checking on the extracted information whether it is identical to the original information sent by user. The algorithms are considered not effective because the receiver does not know the trustworthy of information received. There is also a limitation on the embedding capacity of cover image as not every image can be embedded into cover image.

In this paper, an improved RDWT-based image steganography scheme is proposed to enhance the imperceptibility of image. The related work that uses RDWT and matrix factorization techniques will be discussed in Section II. Section III explains the embedding and extraction process of the proposed work. It shows how secret image will be embedded into cover image by using the double entropy approach. To test the performance of the proposed method, tests for imperceptibility, robustness, false-positive and computational time will be conducted and presented in Section IV. The obtained results will be compared with previous work. Section V concludes this work.

II. Literature Review

A. Frequency Domain Transform Techniques

To be able to achieve good quality and robustness to attacks [2], embedding in transform domain is much more efficient than embedding in spatial domain. The most commonly used frequency-domain transform methods include the Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT) and Redundant Discrete Wavelet Transform (RDWT). Frequency-domain methods are more widely applied as compared to spatial-domain method.

In information hiding schemes, DFT coefficients gives only modest results and is fragile to attacks, especially sensitive to JPEG and MPEG attacks. Capacity and lack of HVS models are also the drawback of DFT [3] whereas DCT based techniques are robust against simple image processing operations but they are hard to implement, takes more computational time and cost and weak against geometric attacks.

Discrete Wavelet Transform (DWT) is a modern technique popularly utilized in digital image processing. The transforms are based on wavelet of limited duration and different frequency. The wavelet transform decomposes the image into three directions. Most of the image energy concentrates at LL band. Hence, embedding in other sub-bands would lower the quality of image. Hiding information in the transform domain is generally more robust [4, 5] and less perceptible.

As compared to DCT as adopted by Lai [6], DWT has more advantages. As it does not suffer from blocking artefacts, it takes less computational time. Therefore, information hiding techniques based on wavelets are more robust [7] against attacks than those based on DCT.

Comparing DWT and DCT, DWT has better energy compaction and presents a sparse time-frequency but there is a major disadvantage of DWT which is the poor directional selectivity and lack of shift variance.

Redundant Discrete Wavelet Transform (RDWT) is a shift invariance property. Let be the input signal and be its reconstructed version. and are low pass and high pass analysis filters while and are corresponding low pass and high pass synthesis filters and are output coefficients at level j. RDWT avoids down and up sampling of coefficients. During image extraction process, DWT produce inaccuracy [8] because of its shift variances property. Many information hiding schemes apply RDWT to overcome the shirt variance problem [9] of DWT. It removes the down-sampling operation from DWT to produce an over-complete representation [10] of the frequency coefficients. Also, as compared to DWT [11], RDWT is more robust. Besides that, RDWT helps to enhance embedding capacity [12] because its sub-bands have the same size of the original image.

B. Matrix Factorization Techniques

There are several matrix factorization methods such as singular value decomposition (SVD), Schur decomposition, QR decomposition, LU decomposition, etc. [13] resulting from solutions of linear equation.

SVD of an matrix A with dimensions m x m is given by

$$A = USV^T \quad (1)$$

With SVD's good stability, applying it to an image does not cause noticeable change [14, 15] on the appearance.

Schur decomposition of a real matrix A results in two matrices U and D such that

$$A = U \times D \times U' \quad (2)$$

It is suggested to be used in digital image processing [16] as it requires less computational cost than SVD.

QR decomposition or QR factorization [17] is decomposition of matrix into an orthogonal matrix and triangular matrix. Any real matrix A can be expressed as:

$$[Q, R] = qr(A) \quad (3)$$

LU factorization is almost similar to QR decomposition. However, QR decomposition has been proven to be more precise for least square problems. LU factorization [12] can only be applied to square matrices whereas QR factorization can be applied to both square and rectangular matrices.

When comparing SVD with QR decomposition, the latter requires less computational complexity. Another feature of QR decomposition is the resistance to some signal processing operations, such as filtering, lossy compression and noise addition. QR decomposition could also solve the major issue [4] of SVD which is its false positive problem.

C. Arnold Transform

Arnold Transform is widely used in image permutation. It is also called the Cat Face transfer, and it is given by.

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} (mod\ N) \quad (4)$$

Arnold transform is often being applied on involved images to improve the security of information hiding schemes.

D. Image Texture Analysis

There are several ways to measure the texture of a digital image. Entropy is the most suitable way to measure the image's texture content. Texture provides measure of properties of an image such as regularity, coarseness and smoothness. An image that is perfectly flat will have an entropy value of zero. Entropy [18] can be defined as the statistical measure of randomness.

According to Shannon's definition, the entropy of a grayscale image is given by the following mathematical relation:

$$E_1 = - \sum_{i=0}^{L-1} p_i \log p_i \qquad (5)$$

If the entropy value is high then it is considered to have more details [19].

E. Performance Evaluation Techniques

To measure the quality of a digital image, PSNR is one of the popular metrics to use. It is done by analyzing the mean squared error value between the Cover and the stego image. The higher the PSNR value [20], the smaller the possibility of visual attack by human eyes. PSNR is being presented as [21].

$$PSNR\ (in\ dB) = 10log_{10}\left(\frac{255^2}{MSE}\right) \qquad (6)$$

and

$$MSE = \frac{\sum_{i=1}^{N}(C_i - C\prime_i)^2}{N} \qquad (7)$$

On the other hand, SSIM is also used as a metric to measure the similarity between two images. It is being done by.

$$\text{SSIM}\ (x, y\ ) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x{}^2 + \mu_y{}^2 + c_1\ )(\sigma_x{}^2 + \sigma_y{}^2 + c_2)} \qquad (8)$$

$$c_1 = (\ k_1 L\ )^2 \qquad (9)$$

$$c_2 = (\ k_2 L\ )^2 \qquad (10)$$

The higher the SSIM index is, the better the visual quality [21] of the stego image with respect to the cover image.

F. Related Work

The spatial domain methods are less robust and use simple embedding process. When the issues of imperceptibility and robustness are concerned, the transform domain techniques are a better option [2], especially when they are being tested against geometric attacks.

Recently, many hybrid transform domain methods are being developed to improve the robustness. While individual transform domain techniques such as DFT, DCT and DWT are good enough to improve the properties as mentioned in 2.6, combining them with Singular Value Decomposition (SVD) will further improve the image quality of the scheme. One of the features of SVD is that it helps to achieve good transparency and robustness.

In 2013, Anumol and Anusudha [22] presented a robust scheme based on discrete wavelet transform and singular value decomposition. The paper chose to use DWT over DFT and DCT as the DFT and DCT are full frame transforms and hence any changes in the transform coefficient affects the entire image. In the embedding process, HL and LH sub-bands are used instead of LL and HH band to improve the robustness and imperceptibility. The proposed method is robust against various types of attacks.

In 2016, Mansi and Vijay [12] proposed an image steganography scheme using QR decomposition and RDWT. Artefacts due to variation in energy distribution caused by shifts in input signal have been solved by RDWT. QR decomposition requires less computational complexity and eliminates false positive issue. The proposed work has shown improvement in terms of robustness, capacity and imperceptibility.

In 2016, Taha et al. [15] evaluated the performance of both RDWT-SVD and DWT-SVD schemes. RDWT is a shift invariance property. Many schemes based on RDWT have been proposed to overcome the shift variant problem of DWT. It removes the down-sampling operation from DWT to produce an over-complete representation of the frequency coefficients. Performing SVD on images is computationally costly. By combining SVD with DWT or RDWT, less computational effort is required to produce better performance. Both schemes showed robust against all attacks, RDWT-SVD is better than DWT-SVD, especially for geometrical attacks.

In 2017, Shaoli Jia et al. [4] introduced a scheme based on DWT and QR Decomposition for colour images. The main advantage of using DWT is that it better takes into account the local image characteristics at different resolution levels which can significantly improve the robustness of hidden message. Arnold transform is applied on the message to ensure security. The work used QR decomposition instead because SVD requires greater computational complexity.

In 2018, Poonam et al. [23] presented a scheme for grayscale images based on DWT-SVD. DWT provided better robustness and visible transparency as compared to DCT and DFT. This paper illustrated an improvement in imperceptibility as well as robustness against attacks.

In 2018, Ferda and Muhammad [24] proposed a block-based RDWT-SVD method using human visual system (HVS) characteristics. This scheme presents an embedding method by examining the coefficients in the first column of U vectors. The proposed scheme can avoid the false-positive problem faced by many other RDWT-SVD schemes during extraction process. The hidden image is scrambled by Arnold transform for security purpose. Compared to existing methods, the method achieved better robustness and imperceptibility under various attacks.

In 2018, Divya and [25] developed a false-positive-free scheme based on shuffled SVD (SSVD) and RDWT. The scheme embeds hidden message on the cover image using Redundant Discrete Wavelet Transform (RDWT) and chaotic mapping. Chaotic mapping is achieved by Shuffled singular value decomposition (SSVD). SSVD enhanced the quality of reconstructed image by breaking an image into a set of ensemble images. The proposed work eliminates the false-positive problem that usually found in other RDWT-SVD methods.

Inspired by above-mentioned previous works, an improved image steganography scheme based on RDWT and QR decomposition is proposed to embed grayscale image into grayscale cover image by analyzing the image texture using entropy.

III. Methodologies

This section explains the detailed embedding and extraction process of the proposed work. The proposed scheme made use of the image texture by calculating entropy values for every image block for both cover and secret images before embedding process (refer to Fig. 1) takes place. Fig. 2 shows the extraction process of proposed method.

The main contribution of the proposed work is the use of entropy in the embedding process. Before embedding takes place, both cover image and secret image will be segmented into non-overlapping blocks with the same block size. Every image block will be computed to produce an entropy value. Then, all values will be sorted in descending order, from the highest entropy value to the lowest entropy value. This applies to both cover and secret image.

Starting from the block with the highest entropy value, RDWT will be applied on the blocks. The LL sub-band of cover image block will then be decomposed by QR decomposition. The secret image information is embedded by modifying the R value of the LL sub-band of cover image. After that, inverse QR decomposition is performed to get the modified LL sub-band. Them, inverse RDWT will be applied to get the modified image block. This process will continue until all secret image information is being embedded. Finally, a stego image will be formed by combining all modified image blocks.

The reason to use RDWT instead of other transform techniques is that, RDWT removes the coefficients' up-sampling and down-sampling that exists in DWT. Also, RDWT can increase robustness and provides more embedding capacity. On the other hand, QR decomposition provides better imperceptibility and avoids false positive issue, as compared to SVD.

Entropy identifies the texture of a digital image. The higher the entropy value of an image block, the more details it contains. By utilizing the use of entropy during embedding process, the cover image block with the highest entropy value will be embedded with the secret image block with the highest entropy value. The minor changes in the particular image block will not cause significant effect on the entropy values. Now, imagine the secret image block with the highest entropy value, which contains more details, is embedded into the cover image block with the lowest entropy. This process will cause more changes in the image texture as compared to the previous case, which will indirectly affect the imperceptibility of the scheme.

The proposed work embeds secret image into cover image by following the steps shown in Fig. 1. By considering the order of entropy values of all image blocks, higher imperceptibility is achieved. As the image block with highest entropy value will remain having the highest entropy value as it is embedded with another image block, which, in this case, the secret image block with the highest entropy value. This will not be accomplished if the process replaces any of the image block with entropy value of different order instead of the same order because the image texture will be changed in a different manner that causes more differences.

Fig. 2 demonstrates the extraction process which extracts secret image information from the image location that was set during the embedding process. There is no need for re-calculation of entropy values for every image blocks.

A. Embedding Process

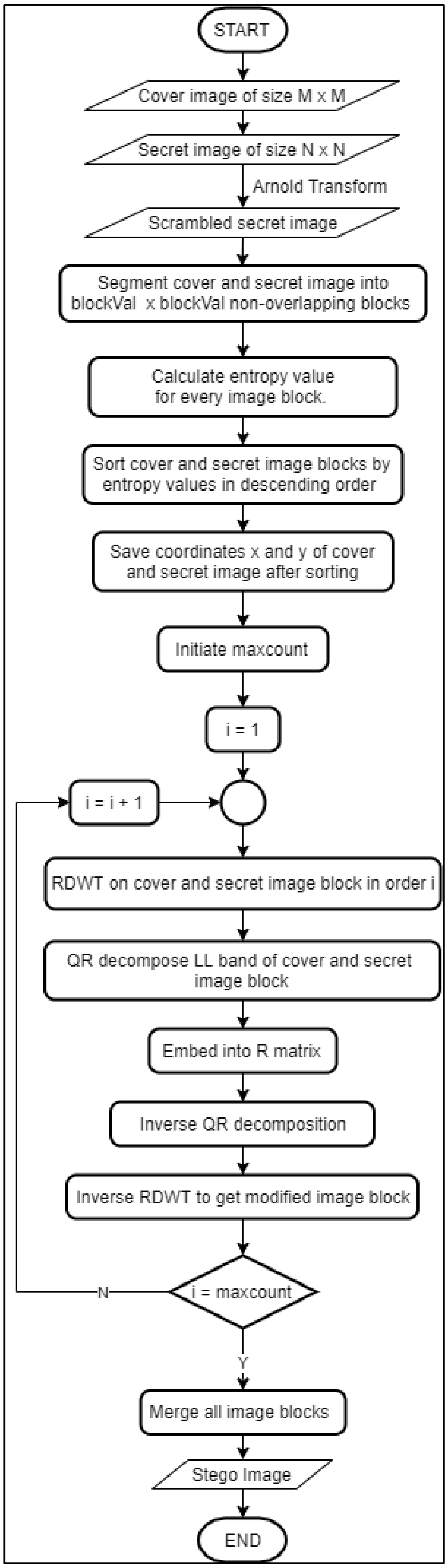

START
Cover image of size M x M
Secret image of size N x N
Arnold Transform
Scrambled secret image
Segment cover and secret image into blockVal x blockVal non-overlapping blocks
Calculate entropy value for every image block.
Sort cover and secret image blocks by entropy values in descending order
Save coordinates x and y of cover and secret image after sorting
Initiate maxcount
i = 1
i = i + 1
RDWT on cover and secret image block in order i
QR decompose LL band of cover and secret image block
Embed into R matrix
Inverse QR decomposition
Inverse RDWT to get modified image block
i = maxcount
N
Y
Merge all image blocks
Stego Image
END

Fig. 1. Embedding Process of Proposed Method.

The embedding algorithm:

```
Input: Cover Image; Secret Image
Pre-processing:
Step 1: Segment cover image, c of size M x M into blocksize x blocksize non-overlapping blocks.
Step 2: Apply Arnold transform on secret image.
    $\mathbf{S}' = \mathbf{arnold}(\mathbf{S}, \mathbf{key})$    (11)
Step 3: Segment secret image, s of size N x N into blocksize x blocksize non-overlapping blocks.
Step 4: Calculate entropy for each cover image blocks.
Step 5: Sort the cover image blocks according to entropy values from highest to lowest.
Step 6: Save coordinates of $x_c$ and $y_c$ after sorting.
Step 7: Calculate entropy for each secret image blocks.
Step 8: Sort the secret image blocks according to entropy values from highest to lowest.
Step 9: Save coordinates of $x_s$ and $y_s$ after sorting.
Secret Image Embedding:
Step 10: Select cover image blocks from the highest entropy value to the lowest.
Step 11: Each selected blocksize x blocksize cover image block is transformed by 1-level RDWT.
    $[\mathbf{LL}_c, \mathbf{LH}_c, \mathbf{HL}_{c,}\mathbf{HH}_c] = \mathbf{rdwt}(\mathbf{cover\ image\ block})$    (12)
Step 12: Select secret image blocks from the highest entropy value to the lowest.
Step 13: Each selected blocksize x blocksize secret image block is transformed by 1-level RDWT.
    $[\mathbf{LL}_s, \mathbf{LH}_s, \mathbf{HL}_{s,}\mathbf{HH}_s] = \mathbf{rdwt}(\mathbf{secret\ image\ block})$    (13)
Step 14: The LL subband of cover image block is decomposed by QR decomposition.
    $[\boldsymbol{Q}_{c,}, \boldsymbol{R}_c] = \mathbf{qr}(\mathbf{LL}_c)$    (14)
Step 15: The LL subband of secret image block is decomposed by QR decomposition.
    $[\boldsymbol{Q}_{s,}, \boldsymbol{R}_s] = \mathbf{qr}(\mathbf{LL}_s)$    (15)
Step 16: The secret image is embedded by modifying the R value of the $LL_c$.
    $\boldsymbol{R}_{st} = \boldsymbol{R}_c + \propto \boldsymbol{R}_s$    (16)
Step 17: Perform inverse QR decomposition to get the modified $LL_c$ to form $LL_{st}$.
    $\mathbf{LL}_{st} = \boldsymbol{Q}_c \times \boldsymbol{R}_{st}$    (17)
Step 18: Apply inverse RDWT to get the modified image block
    $\mathbf{Modified\ image\ block} = \boldsymbol{RDWT}^{-1}(\mathbf{LL}_{st}, \mathbf{LH}_c, \mathbf{HL}_{c,}\mathbf{HH}_c)$
    (18)
Step 19: Repeat step 10-18 until all secret image blocks have been embedded.
Post-processing:
Step 20: Merge all modified image blocks to form Stego Image.
Output: Stego Image
```

B. Extraction Process

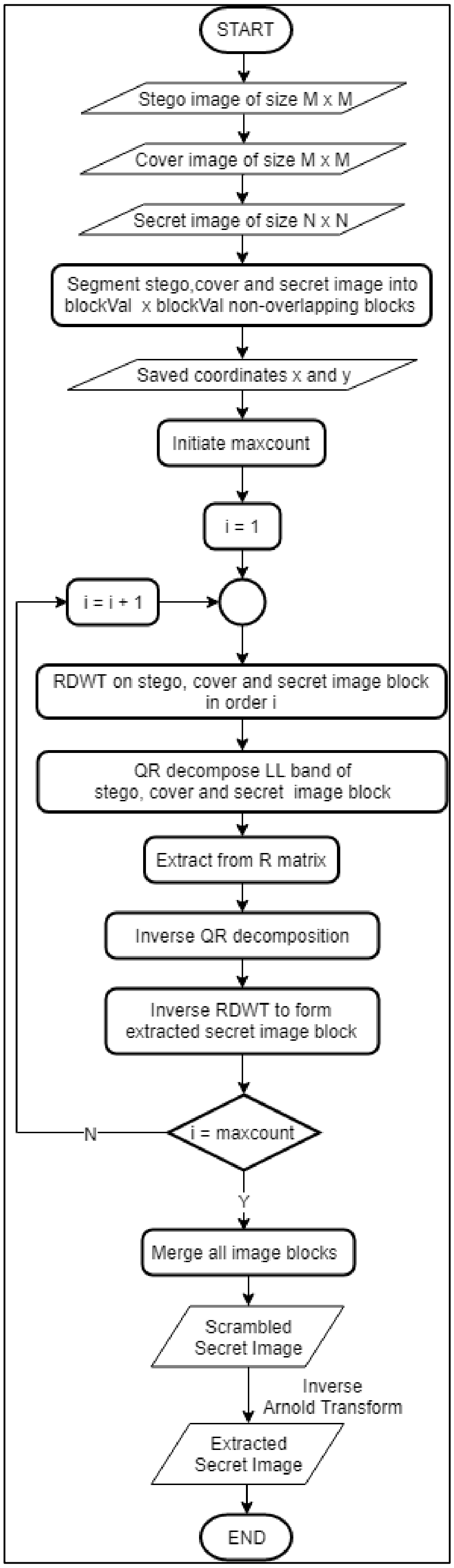

START
Stego image of size M x M
Cover image of size M x M
Secret image of size N x N
Segment stego,cover and secret image into blockVal x blockVal non-overlapping blocks
Saved coordinates x and y
Initiate maxcount
i = 1
i = i + 1
RDWT on stego, cover and secret image block in order i
QR decompose LL band of stego, cover and secret image block
Extract from R matrix
Inverse QR decomposition
Inverse RDWT to form extracted secret image block
i = maxcount
N
Y
Merge all image blocks
Scrambled Secret Image
Inverse Arnold Transform
Extracted Secret Image
END

Fig. 2. Extraction Process of Proposed Method.

The extraction algorithm:

Input: Stego Image; Cover Image; Secret Image

Pre-processing:

Step 1: Segment stego image , st of size M x M into blocksize x blocksize non-overlapping blocks.

Step 2: Segment cover image, c of size M x M into blocksize x blocksize non-overlapping blocks.

Step 3: Segment secret image, s of size N x N into blocksize x blocksize non-overlapping blocks.

Secret Image Extraction:

Step 4: Select stego image blocks from the highest entropy value to the lowest according to the x_c and y_c coordinates obtained.

Step 5: Each selected blocksize x blocksize stego image block is transformed by 1-level RDWT.

$$[\mathbf{LL}_{st}, \mathbf{LH}_{st}, \mathbf{HL}_{st,} \mathbf{HH}_{st}] = \mathbf{rdwt(stego\ image\ block)} \quad (19)$$

Step 6: Select cover image blocks from the highest entropy value to the lowest according to the x_c and y_c coordinates obtained.

Step 7: Each selected blocksize x blocksize cover image block is transformed by 1-level RDWT.

$$[\mathbf{LL}_c, \mathbf{LH}_c, \mathbf{HL}_{c,} \mathbf{HH}_c] = \mathbf{rdwt(cover\ image\ block)} \quad (20)$$

Step 8: Select secret image blocks from the highest entropy value to the lowest according to the x_s and y_s coordinates obtained.

Step 9: Each selected blocksize x blocksize secret image block is transformed by 1-level RDWT.

$$[\mathbf{LL}_s, \mathbf{LH}_s, \mathbf{HL}_{s,} \mathbf{HH}_s] = \mathbf{rdwt(secret\ image\ block)} \quad (21)$$

Step 10: The LL subband of stego image block is decomposed by QR decomposition.

$$[\boldsymbol{Q}_{st,}, \boldsymbol{R}_{st}] = \mathbf{qr}(\mathbf{LL}_{st}) \quad (22)$$

Step 11: The LL subband of cover image block is decomposed by QR decomposition.

$$[\boldsymbol{Q}_{c,}, \boldsymbol{R}_{c}] = \mathbf{qr}(\mathbf{LL}_{c}) \quad (23)$$

Step 12: The LL subband of secret image block is decomposed by QR decomposition.

$$[\boldsymbol{Q}_{s,}, \boldsymbol{R}_{s}] = \mathbf{qr}(\mathbf{LL}_{s}) \quad (24)$$

Step 13: The secret image is extracted by

$$\boldsymbol{R}_{s2} = (\boldsymbol{R}_{st} - \boldsymbol{R}_c) \,/ \propto \quad (25)$$

Step 14: Perform inverse QR decomposition to form LL_{s2}.

$$\mathbf{LL}_{s2} = \boldsymbol{Q}_s \times \boldsymbol{R}_{s2} \quad (26)$$

Step 15: Apply inverse RDWT to get the extracted secret image block.

$$\mathbf{Extracted\ secret\ image\ block} = \boldsymbol{RDWT}^{-1}(\mathbf{LL}_{s2}, \mathbf{LH}_s, \mathbf{HL}_{s,} \mathbf{HH}_s) \quad (27)$$

Step 16: Repeat step 4-15 until all secret image blocks have been extracted.

Post-processing:

Step 17: Merge all extracted image blocks to form image S2.

Step 18: Apply inverse Arnold transform on image S2 to get secret image S_{ext}.

$$\boldsymbol{S}_{ext} = \mathbf{inarnold(S2, key)} \quad (28)$$

Output: Extracted Secret Image

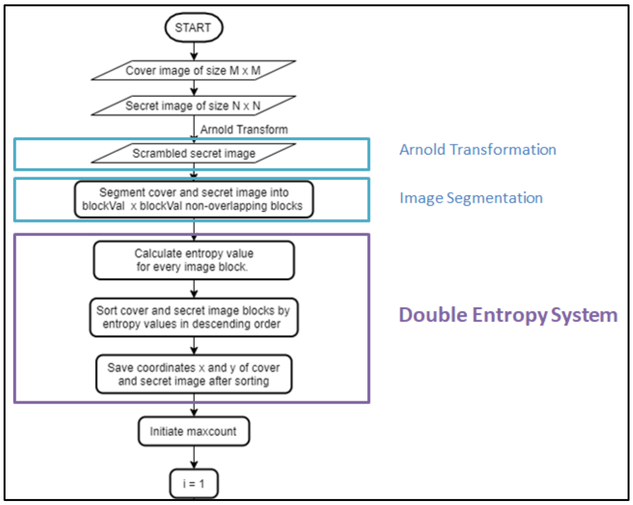

START
Cover image of size M x M
Secret image of size N x N
Arnold Transform
Scrambled secret image
Arnold Transformation
Segment cover and secret image into blockVal x blockVal non-overlapping blocks
Image Segmentation
Calculate entropy value for every image block.
Sort cover and secret image blocks by entropy values in descending order
Save coordinates x and y of cover and secret image after sorting
Double Entropy System
Initiate maxcount
i = 1

Fig. 3. Part of Embedding Process.

C. Double Entropy System

The double entropy system is what differentiates this proposed scheme from other steganography schemes. It happens right before embedding process takes place as shown in Fig. 3.

The double entropy system considers both entropy values of cover image blocks and secret images blocks before embedding process begins. By making use of the human visual system, cover and secret image have been segmented into blocks of equal size and each image block has an entropy value. All entropy values are then being sorted in descending order.

When embedding process starts, secret image block with the highest entropy value will be embedded into the cover image block with the highest entropy value. The embedding process continues until all secret image blocks have been embedded into cover image according to their corresponding entropy values, from the highest value to the lowest value.

By initiating embedding process according to the entropy values of both cover and secret image blocks, the block that higher entropy value, hence, more details will be embedded into another block that has more details. Through this process, the cover image block that originally has the high level of detail will remained as it is but it now contains the secret image information. The block remains having high entropy values and has least impact on human visual system as it is harder to notice the difference before and after embedding as compared to block with lower entropy value. Proposed work utilizes the double entropy system that applies entropy calculation on both cover and secret image in order to find out the areas that are more suitable for embedding to achieve a better-quality stego image. When there are changes or modification happens in that particular area, HVS will less likely to notice the difference because of its low sensitivity towards the area. It is to believe that the imperceptibility of proposed scheme will be improved through this approach.

IV. Results and Discussion

To demonstrate the effect of different aspects of proposed scheme, different experiments have been carried out including the imperceptibility test, false positive test and computational time evaluation.

The images used in these experiments are lena, baboon, peppers, lake, house, jetplane, livingroom, pirate, bridge, boat, cameraman and barbara as demonstrated in Table IX. The cover image size is set to be 512x512. The sizes of secret image are of 32x32, 64x64, 128x128, 256x256 and 512x512 in order to compare how image size affects the imperceptibility of stego image.

A. Imperceptibility Test

The proposed work consists of three techniques, which include RDWT, QR decomposition and double entropy. To show how adding double entropy into the scheme helps with achieving better imperceptibility, two different experiments have been carried out. There will be one experiment using double entropy in the embedding process and the other one will eliminate double entropy during the embedding process.

Every experiment uses different combination of cover image and secret image to generate PSNR and SSIM values. In order to achieve a fair comparison of imperceptibility, the cover and secret images used in proposed scheme will be the same as the work being compared with. The experiments will be carried out by varying the size of segmented block and the size of secret image in order to show how they affect the imperceptibility of the scheme.

The average PSNR and SSIM values for each experiment are presented in Table I, Table II, Table III, Table IV and Table V. The average PSNR values of all experiments are being presented in Table VI.

From the results above, t is shown that proposed work that uses RDWT, QR decomposition and double entropy has the highest PSNR value compared to the other two experiments. On the other hand, the experiment that uses RDWT and QR decomposition without including double entropy gives the second highest PSNR value while the one that uses only RDWT and double entropy without QR decomposition generates the lowest PSNR value.

By analysing on the average PSNR values for different secret image sizes, it is believed that the smaller the secret image size, the higher the PSNR value. This is because of the smaller amount of information being embedded into cover image, hence, smaller modification made on the cover image.

In order to achieve a fair comparison of imperceptibility between proposed scheme and existing work, the cover and secret images used will be the same as the work being compared with.

From the comparison of PSNR and SSIM values presented in Table VII and Table VIII, it is shown that proposed scheme performed better in terms of imperceptibility as compared to existing work due to the utilization of double entropy system in embedding process.

TABLE. I. Average PSNR (without Double Entropy)

Secret Image Size	**Block Size**			
	4 x 4	**8 x 8**	**16 x 16**	**32 x 32**
32 x 32	74.9414	74.8407	74.7871	-
64 x 64	68.5213	68.5471	68.5892	68.6398
128 x 128	62.3082	62.3500	62.4002	62.7121
256 x 256	56.1674	56.2770	56.2964	56.3785
512 x 512	50.1361	50.1827	50.2526	50.3078

TABLE. II. Average SSIM (without Double Entropy)

Secret Image Size	**Block Size**			
	4 x 4	**8 x 8**	**16 x 16**	**32 x 32**
32 x 32	0.9999	1.0000	1.0000	-
64 x 64	0.9997	0.9998	0.9999	0.9999
128 x 128	0.9991	0.9994	0.9996	1.0830
256 x 256	0.9976	0.9985	0.9989	0.9991
512 x 512	0.9977	0.9983	0.9985	0.9921

TABLE. III. Average PSNR (with Double Entropy)

Secret Image Size	**Block Size**			
	4 x 4	**8 x 8**	**16 x 16**	**32 x 32**
32 x 32	74.9260	74.9106	74.9695	-
64 x 64	68.5912	68.6517	68.6974	68.8272
128 x 128	62.3726	62.4577	62.5343	62.6374
256 x 256	56.2711	56.3604	56.4219	56.4909
512 x 512	50.1862	50.2788	50.3361	50.3986

TABLE. IV. Average SSIM (with Double Entropy)

Secret Image Size	**Block Size**			
	4 x 4	**8 x 8**	**16 x 16**	**32 x 32**
32 x 32	1.0000	1.0000	1.0000	-
64 x 64	0.9999	1.0000	1.0000	1.0000
128 x 128	0.9999	0.9999	0.9998	0.9998
256 x 256	0.9997	0.9996	0.9995	0.9995
512 x 512	0.9991	0.9989	0.9987	0.9985

TABLE. V. AVERAGE PSNR (WITHOUT QR DECOMPOSITION)

Secret Image Size	Block Size			
	4 x 4	8 x 8	16 x 16	32 x 32
32 x 32	74.8736	74.7302	74.6736	-
64 x 64	68.5102	68.4552	68.4295	68.4037
128 x 128	62.2878	62.2651	62.2560	62.2435
256 x 256	56.1576	56.1444	56.1386	56.1312
512 x 512	50.0773	50.0644	50.0609	50.0575

TABLE. VI. AVERAGE PSNR VALUES BETWEEN 3 EXPERIMENTS

Secret Image Size	Experiments		
	Without Double Entropy	With Double Entropy	Without QR Decomposition
32 x 32	74.8564	**74.9354**	74.7591
64 x 64	68.5744	**68.6919**	68.4497
128 x 128	62.4426	**62.5005**	62.2631
256 x 256	56.2798	**56.3861**	56.1430
512 x 512	50.2198	**50.2999**	50.0650

TABLE. VII. COMPARISON OF IMPERCEPTIBILITY WITH DIVYA AND RANJAN'S WORK BY PSNR VALUES

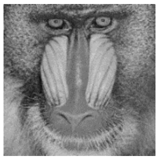
Secret Images	Cover Image	Divya and Ranjan's	Proposed
 and 的		54.3157	60.3666
		54.0472	60.3674
		58.3684	60.3980

TABLE. VIII. COMPARISON OF IMPERCEPTIBILITY WITH FERDA'S WORK BY SSIM VALUES

Secret Image	Cover Image	Ferda's	Proposed
福		 SSIM = 0.9965	 SSIM = **0.9999**
福		 SSIM = 0.9968	 SSIM = **0.9998**

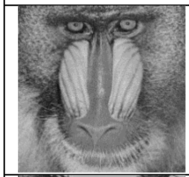
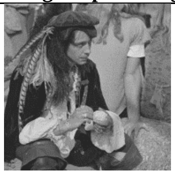
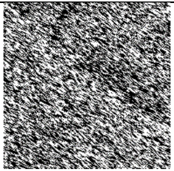
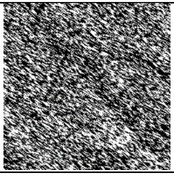
B. False Positive Test

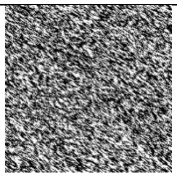
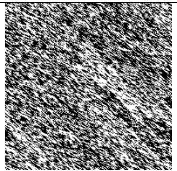
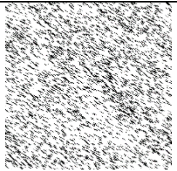
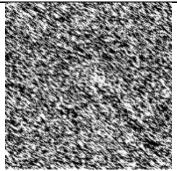
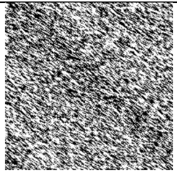
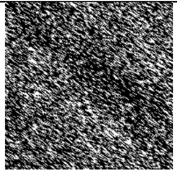
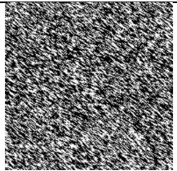
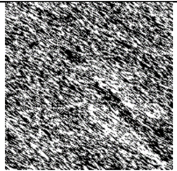
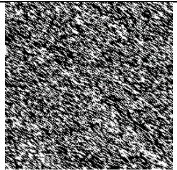
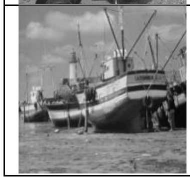
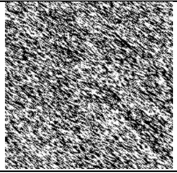
A false positive issue occurs when the image extracted from an unauthorized or arbitrary image shows a visual trace of the owner's original embedded image. This test is conducted using the proposed RDWT-QR scheme with double entropy system. The original embedding and extraction process are done by using Baboon as the cover image and Pirate as the embedded secret image.

Table IX shows the extracted images and respective NC values using the original stego image (Baboon) and other test images.

From the results shown in Table IX, it showed that proposed work can extract secret image from the correct stego image with high NC value (i.e. 0.8889). By changing the stego image to 12 other test images, the extracted image is meaningless and does not ensemble the original embedded secret image. The NC values obtained is very low, ranging from -0.0066 to 0.0127. Therefore, it is proved that the false positive issue that normally occurred in other steganography schemes can be avoided by using proposed algorithm.

TABLE. IX. FALSE POSITIVE TEST RESULTS FOR PROPOSED METHOD

Test Image(s)	Extracted Secret Image using Proposed Algorithm	NC Values
		0.8889
		-0.0065
		0.0094

		0.0026
		-0.0033
		0.0108
		-0.0112
		0.0085
		0.0091
		-0.0066
		0.0127
		0.0110
		-0.0008

C. *Computational Time*

Table X shows the total embedding time for secret image of different sizes with varying block size whereas Table XI presents the total extraction time for secret image of different sizes with varying block size.

The time taken to embed secret image is slightly longer than the extraction process as shown in the comparison presented in Table XII. The reason is that the embedding of secret image takes more time during the selection of position to embed after calculating entropy values for each segmented image block. This process has been shortened during the extraction process as the position to extract secret information has already been set because of the saved coordinates of x and y.

Table XIII shows that proposed work spent less computational time on embedding process than Ferda's [24] and Divya and Ranjan's [25] work. For the extraction process, proposed work spent a slightly longer time compared to Divya and Ranjan's and 24 seconds faster than Ferda's. Due to the lower complexity of the algorithm, the time taken to execute the embedding and extraction process is shorter.

TABLE. X. EMBEDDING TIME FOR VARYING SECRET IMAGE SIZE WITH DIFFERENT BLOCK SIZE (IN SECONDS)

Secret Image Size	**Block Size**			
	4 x 4	**8 x 8**	**16 x 16**	**32 x 32**
32 x 32	0.78	0.34	1.11	-
64 x 64	3.59	0.69	1.22	0.30
128 x 128	9.33	3.84	1.66	0.50
256 x 256	20.80	9.83	4.61	1.23
512 x 512	61.72	24.52	10.38	5.02

TABLE. XI. EXTRACTION TIME FOR VARYING SECRET IMAGE SIZE WITH DIFFERENT BLOCK SIZE (IN SECONDS)

Secret Image Size	**Block Size**			
	4 x 4	**8 x 8**	**16 x 16**	**32 x 32**
32 x 32	0.67	0.14	0.08	-
64 x 64	3.61	0.80	0.19	0.16
128 x 128	7.42	3.70	1.47	0.38
256 x 256	19.05	8.08	4.17	1.15
512 x 512	60.32	22.74	8.83	4.81

TABLE. XII. AVERAGE COMPUTATION TIME FOR PROPOSED WORK (IN SECONDS)

Secret Image Size	**Embedding**	**Extraction**
32 x 32	0.74	0.30
64 x 64	1.45	1.19
128 x 128	3.83	3.24
256 x 256	9.12	8.11
512 x 512	25.41	24.18

TABLE. XIII. COMPARISON OF COMPUTATION TIME

Embedding (in seconds)			**Extraction (in seconds)**		
Ferda's	**Ranjan's**	**Proposed**	**Ferda's**	**Ranjan's**	**Proposed**
86.2969	0.79	**0.72**	25.3125	**0.22**	1.14

V. Conclusion

To test the effectiveness of the proposed scheme, several tests have been conducted. These include the imperceptibility test, false-positive test and computational time taken for both embedding and extraction process.

Test results have shown that proposed scheme has higher imperceptibility and provides image with better quality as compared to previous work by applying the double entropy system.

In terms of computational time and cost, the time taken to embed secret image into cover image and the time taken to extract secret image from stego image are way faster than compared work. Due to the less complex execution of the proposed algorithm, it has reduced the time taken to complete both embedding and extraction process.

On the other hand, the proposed scheme provides a certain level of security by applying Arnold transform on secret image before embedding. It also eliminates the occurrence of false-positive errors, which has been found in many other existing works. Furthermore, RDWT allows embedding of the same-sized secret image into cover image as compared to DWT that only offers half the embedding capacity of RDWT. It also solves the shirt variance problem that caused by DWT to avoid inaccuracy during extraction process.

As a conclusion, the comparison results have shown that proposed work enhanced the imperceptibility of the scheme and eliminates the false-positive issue. Also, proposed work took shorter time to execute embedding and extraction process.

Acknowledgment

The authors sincerely thank Universiti Malaysia Pahang, Malaysia, for providing financial support for this work through UMP Postgraduate Research Grants Scheme (RDU1703280).

References

[1] Thiyagarajan, P.; Aghila, G.; Prasanna, Venkatesan V., "Stego-Image Generator (SIG) – Building Steganography Image Database," CDBR-SSE Lab Department of Computer Science, Pondicherry University, Puducherry 605014, 2012.

[2] Prerna Gupta, Girish Parmar, "Image Watermarking using IWT-SVD and its Comparative Analysis with DWT-SVD," in International Conference on Computer, Communications and Electronics (Comptelix), 2017.

[3] Serdean C.V., Tomlinson M., Wade G.J., Ambroze A.M., "Protecting intellectual rights: Digital watermaking in the wavelet domain," in Trends and Recent Achievements in Information Technology, 2002. K. Elissa, "Title of paper if known," unpublished.

[4] Shaoli Jia, Qingpo Zhou and Hong Zhou, "A Novel Color Image Watermarking Scheme Based on DWT and QR Decomposition," Journal of Applied Science and Engineering, pp. 193-200, 2017.

[5] L.-Y. Hsu, H.-T. Hu, "Robust blind image watermarking using crisscross inter-block prediction in the DCT domain," J. Vis. Commun. Image R., pp. 33-47, 2017.

[6] Lai, C.C, "An improved SVD-based watermarking scheme using human visual," Optical Commununication, pp. 938-944, 2011.

[7] Serdean C.V., Tomlinson M., Wade G.J., Ambroze A.M., "Protecting intellectual rights: Digital watermaking in the wavelet domain," in Trends and Recent Achievements in Information Technology, 2002.

[8] Bradley, A.P., "Shift-invariance in the discrete wavelet transform," in Proc. VIIth Digital Image Computing: Techniques and Applications, 2003.

[9] Hien T.D., Nakao Z., Chen Y.W., "RDWT domain watermarking based on independent component analysis extraction," Advance Software Computing, pp. 401-414, 2006.

[10] Nasrin M. Makbol, Bee Ee Khoo, Taha H. Rassem, "Block-based discrete wavelet transform singular value decomposition image watermarking scheme using human visual system characteristics," IET Image Process, pp. 34-52, 2016.

[11] F. JE., "The redundant discrete wavelet transform and additive noise," Signal Processing Letters, pp. 629-632, 2005.

[12] M.S. Subhedar, V.H. Mankar, "Image steganography using redundant discrete wavelet transform and QR factorization," Computers and Electrical Engineering, pp. 406-422, 2016.

[13] Q. Su et al., "Embedding color watermarks in color images based on Schur decomposition," Optics Communications 285, pp. 1792-1802, 2012.

[14] Sunil et al., "An Improved Image Steganography based on 2-DWT-FFT-SVD on YCBCR Color Space," in International Conference on Trends in Electronics and Informatics, 2017.

[15] Taha H. Rassem, Nasrin M. Makbol, and Bee Ee Khoo, "Performance evaluation of RDWT-SVD and DWT-SVD watermarking schemes," in International Conference on Advanced Science, Engineering and Technology (ICASET), 2016.

[16] G.H. Golub, C.F. Van Loan, "Matrix Computations," Johns Hopkins University Press, Baltimore, 1989.

[17] C.-J. Ahn, "Ahn CJ . Parallel detection algorithm using multiple QR decompositions with permuted channel matrix for SDM/OFDM.," IEEE Transactions on Vehicular Technology, pp. 2578-2582, 2008.

[18] Manoj Kumar, Gursewak Singh, "Block based Image Steganography using Entropy with LSB and 2-bit Identical Approach," International Journal of Computer Applications (0975 – 8887), 2017.

[19] Swati Bhargava and Manish Mukhija, "Hide Image and Text Using LSB, DWT and RSA Based on Image Steganography," ICTACT JOURNAL ON IMAGE AND VIDEO PROCESSING, pp. 1940-1946, 2019.

[20] Hossain et al., "Variable Rate Steganography in Gray Scale Digital Images Using Neighborhood Pixel Iriformation," in Proceedings of 2009 12th International Conference on Computer and Information Technology (ICCIT 2009), Dhaka, Bangladesh, 2009.

[21] I.J. Kadhim, P. Premaratne and P.J. Vial et al., "Comprehensive survey of image steganography: Techniques, Evaluations, and trends in future research," Neurocomputing, pp. 299-326, 2019.

[22] Anumol Joseph, K. Anusudha, "Robust watermarking based on DWT SVD," International Journal of Signal & Image Processing, 2013.

[23] Poonam, Shaifali M.Arora, "A DWT-SVD based Robust Digital Watermarking for Digital Images," in International Conference on Computational Intelligence and Data Science (ICCIDS), 2018.

[24] F. Ernawan, M. N. Kabir, "A block-based RDWT-SVD image watermarkingmethod using human visual system characteristics," The Visual Computer, 2018.

[25] J L Divya Shivani, Ranjan K. Senapati, "False-positive-free, Robust and Blind Watermarking Scheme based on Shuffled SVD and RDWT," Journal of Advanced Research in Dynamical and Control Systems, 2018

Spectrum Occupancy Measurement of Cellular Spectrum and Smart Network Sharing in Pakistan

Aftab Ahmed Mirani[1], Sajjad Ali Memon[2], Saqib Hussain[3], Muhammad Aamir Panhwar[4], Syed Rizwan Ali Shah[5]
Department of Telecommunication Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan[1, 2, 3, 5]
School of Electronic Engineering, Beijing University of Posts and Telecommunications, China[4]

Abstract—**In wireless communication, the radio spectrum is a very rare and precious resource that has currently become a major problem to efficiently exploit the underutilized band of the static allocated licensed band. Recently, the cognitive radio (CR) has emerged as a promising technology to overcome the spectrum crisis, in which, the licensed band can be utilized by the unlicensed user until and unless it does not affect the transmission of the licensed band. In this paper, the spectrum occupancy of three bands i.e. GSM 900, 1800 and 2100 bands have been measured through spectrum analyzer in the indoor and outdoor environment. The measured results of all the three bands have been calculated through MATLAB against the power spectral density versus frequency plots. Results have shown that the majority of the licensed band is underutilized. Therefore, the CR can play a pivotal role to efficiently utilize the unused spectrum and to overcome the cellular wireless spectrum crisis in Pakistan. The second part of this paper deals with the emerging concept of network sharing among mobile operators and its impacts on cost. Network sharing become the standard among mobile operators worldwide and so in Pakistan. Capital (CAPEX) as well as operational (OPEX) expenditure and rapid advancement in technology encouraged all operators to go for sharing business models. In Pakistan, all four mobile operators Jazz, Telenor, Zong, and Ufone are actively adopting this model to maintain EBITDA (earnings before interest, taxes, depreciation, and amortization). Mainly there are two types of network sharing, Passive infrastructure sharing, and active resource sharing. Passive network sharing is widely used in Pakistan among operators.**

Keywords—Cognitive radio; spectrum occupancy; cellular networks; spectrum analyzer; mobile network operators; sharing models; passive infrastructure sharing

I. Introduction

With the rapid evolution in wireless technological services, the demand to provide ubiquitous high data rates are increasing exponentially. This leads to utilizing the spectrum resource efficiently [1]-[2]. Thus, researchers have come up with an idea to utilize the unused spectrum through smart sensing of available spectrum through various technologies such as cognitive radio (CR). The CR can effectively deal with the spectrum scarcity problems also; it permits secondary users (SUs) to use the unoccupied spectrum without affecting the transmission of primary users (PUs)[3]-[4]. In the meantime, the Federal Communication Commission (FCC) investigated the spectrum occupancy and found that a major portion of the available spectrum is unused due to the static allocation technique [5][6][7]. The CR technology is viewed as an efficient solution to mitigate the spectrum scarcity problem, which is expected to be used in 5G technologies. A CR is a smart communication system that senses the unused spectrum in its surrounding environment and adjusts its spectrum utilization according to its environment [8].

Dynamic Spectrum Access (DSA) is considered one of the predominant approaches to address the challenges of the unutilized spectrum. In DSA, the spectrum sensing is the first step to utilize the spectrum [9]. The primary user is preferred to use the band or channel over a secondary user whereby, a secondary user utilizes the available band in the absence of a primary user [10]. Different frequency bands have been considered in order to be utilized with a DSA. Out of which Very High Frequency (VHF) and Ultra High Frequency (UHF) bands are the most favorable ones. These bands are traditionally used for different applications such as for TV broadcasting, as an outcome, empty channels or frequency in these traditional bands are often called TV White Spaces (TVWSs) [11]. Several international regulatory bodies have proposed solutions to utilize spectral opportunities in TVWSs. One of the most important IEEE standards is IEEE 802.22 which was approved by IEEE in 2011 [12].

The number of subscribers of cellular networks worldwide is increasing at a very fast pace; with almost 164 million subscribers in Pakistan alone. There is an incessant requirement to build Cellular Network infrastructure to meet the demands of the ever-increasing number of mobile users and provide better services, mainly consists of a voice call, short message services (SMS) and data services. The main infrastructure is the Global System of Mobile Telecommunications (GSM) Base station that consists of the Base Transceiver System (BTS). In order to enhance footprint across the country at a rapid pace became challenging for every operator due to high Capital expenditure (CAPEX), increasing operational expenses (OPEX), government regulatory bodies' approvals, taxations, security issues, and infrastructure expenses, etc. To mitigate these challenges, the concept of network sharing emerged and opened new doors of opportunities for Mobile network operators [13]. Currently, there are four mobile operators Jazz, Telenor, Zong, and Ufone; serving the Pakistan Telecom industry and facing the same challenges [14]. This paper provides different scenarios of Network sharing among Mobile operators in Pakistan and the impact of Infrastructure sharing.

In the same context, whenever any operator intends for rollout, it involves huge investment and the necessity to recover it by imposing big charges on subscribers. This leads to less affordability and eventually discourages mobile service providers to shift towards the next generations in this competitive market [15]. More or less, this problem is addressed with the help of network sharing business model in Pakistan. There are several pros of cellular sharing, increased revenue generation, growth in the number of subscribers and market penetration, fast network roll-out, avoids environmental hazardous, maintained control over mobility and above all the significant reduction in cost [16]. However, there are a few cons also, operators may lose independence over network strategy, the hidden cost may affect the revenue of a particular cluster and it is impossible to foresee everything that has to be the part of this agreement. Apparently, the advantages overweigh the disadvantages and it appeals to every mobile operator in the country to adopt the sharing business models for their future strategy.

In this paper, the spectrum occupancy measurement of all the GSM cellular operators of Pakistan in measured through spectrum analyzer in three different bands i.e. 900, 1800 and 2100 bands. The measurements are calculated at different locations of Pakistan including the Telecommunication Department of Mehran University of Engineering & Technology, Jamshoro on several days during university lab hours. The results are carried out in the form of power spectral density (PSD) for different bands of cellular operators working in Pakistan and its bandwidth utilization percentage. The measured results are very helpful for regulatory authorities such as Pakistan Telecommunication Authority (PTA) to revise spectrum occupancy policies and to promote strong competition among cellular operators and cheap rates because every operator intends to have less CAPEX & OPEX.

The formation of this paper is as follows. Section I contains the introduction. The measurement setup and occupancy measurement method are explained in Section II. The measured results are described in Section III. Section IV contains the concluding remarks and future work.

II. Measurement Setup and Method

A. Measurement Setup

The spectrum occupancy measurements have been carried out at various locations of the Sindh region such as Nooriabad, Hyderabad, Autobahn Road, Qasimabad, Site Area and TD of MUET Jamshoro using Rohde & Schwarz FSH6 handheld spectrum analyzer (SA). The SA can operate in the range of 100 KHz to 6 GHz. The spectrum occupancy measurement setup has been demonstrated in Fig. 1, which shows that the GSM 900 antenna is attached to SA via optical fiber cable and the SA is connected to a laptop via optical USB cable. Moreover, the antenna configuration is shown in Table I. The measurements have been carried out for a month at different places of Sindh Region, Pakistan and each measurement has been carried out for about 30 minutes. The Rhode & Schwarz FSH view and MATLAB have been utilized to take records and plot them, respectively.

B. Cellular Operators of Pakistan

There are four GSM cellular mobile service providers working in Pakistan i.e. Telenor, Pakistan Mobile Communication Limited (PMCL/JAZZ), Ufone and Zong. These four cellular operators are using different techniques such as GSM, WCDMA, LTE and LTE-A services to provide its customers ubiquitous high-quality voice and data services[17]. The general methodology is illustrated in the flowchart given below in Fig. 2. The SA settings for GSM service providers and GSM 900, DCS 1800 and 2100 bands in Pakistan are illustrated in Table II.

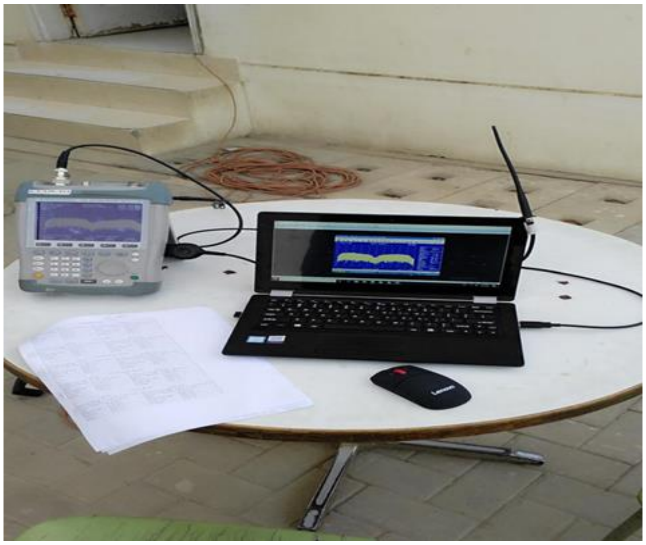


Fig 1. Spectrum Occupancy Measurement Setup.

TABLE I. Antenna Configuration

Electrical Configuration	
Antenna model	AMXT-900-3
Frequency (MHz)	824-960
Bandwidth (MHz)	136
Gain (dBi)	3
VSWR	≤ 2
Impedance	50 Ω
Polarization	Vertical
Maximum input power (W)	50
Input connector type	SMA male
Mechanical Configuration	
Radome Color	Black
Antenna height (mm)	210
Antenna weight (g)	20
Operating temperature (°c)	-40 to 60

TABLE II. THE SA SETTINGS FOR GSM SERVICE PROVIDERS GSM 900 BAND IN PAKISTAN

SA parameters		*Frequency range (MHz)*	*Center frequency (MHz)*	*Frequency span (MHz)*	*Resolution Bandwidth (KHz)*	*Video Bandwidth (KHz)*	*Sweep time (ms)*	*Number of frequency measurement points*
GSM Service Providers								
Mobilink	*Uplink*	*907.3- 914.9*	*911.1*	*7.6*	*200*	*200*	*250*	*301*
	Downlink	*952.3- 959.9*	*956.1*	*7.6*	*200*	*200*	*250*	*301*
Ufone	*Uplink*	*894.9-902.5*	*898.7*	*7.6*	*200*	*200*	*250*	*301*
	Downlink	*939.9-947.5*	*943.7*	*7.6*	*200*	*200*	*250*	*301*
Telenor	*Uplink*	*902.5- 907.3*	*904.9*	*4.8*	*200*	*200*	*250*	*301*
	Downlink	*947.5- 952.3*	*949.9*	*4.8*	*200*	*200*	*250*	*301*
Warid	*Uplink*	*890.1- 894.9*	*892.5*	*4.8*	*200*	*200*	*250*	*301*
	Downlink	*935.1- 939.9*	*937.5*	*4.8*	*200*	*200*	*250*	*301*
Zong	*Uplink*	*882.5- 890.1*	*886.3*	*7.6*	*200*	*200*	*250*	*301*
	Downlink	*927.5- 935.1*	*931.3*	*7.6*	*200*	*200*	*250*	*301*
GSM 900	*Uplink*	*890-915*	*902.5*	*25*	*200*	*200*	*250*	*301*
	Downlink	*935- 960*	*947.5*	*25*	*200*	*200*	*250*	*301*
DCS 1800	*Uplink*	*1718.9-1781.1*	*1750*	*62.2*	*200*	*200*	*250*	*301*
	Downlink	*1813.9-1876.1*	*1845*	*62.2*	*200*	*200*	*250*	*301*
2100	*Uplink*	*1920-1950*	*1935*	*30*	*200*	*200*	*250*	*301*
	Downlink	*2110-2140*	*2125*	*30*	*200*	*200*	*250*	*301*

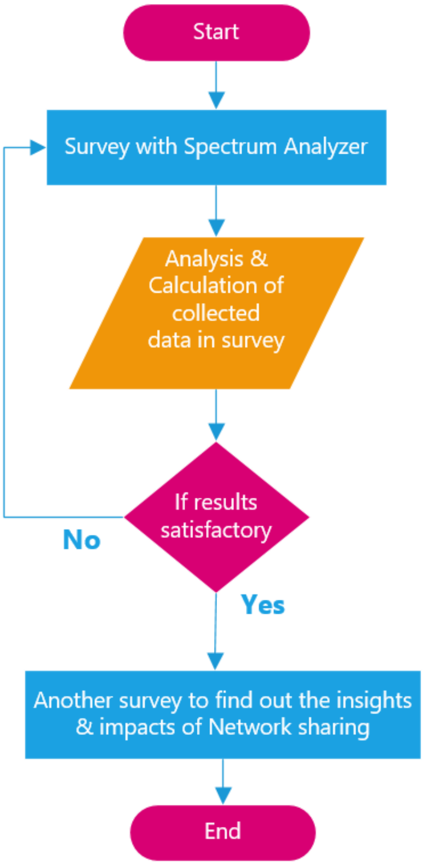

Start
Survey with Spectrum Analyzer
Analysis & Calculation of collected data in survey
If results satisfactory
No
Yes
Another survey to find out the insights & impacts of Network sharing
End

Fig 2. Flowchart of Research Methodology.

III. OCCUPANCY RESULTS AND DISCUSSIONS

The results of the SA measurement are represented in the average power spectral density (APSD) versus frequency plots in the MATLAB. The APSD is measured by averaging all the values gathered during a 30 days' measurement campaign for different bands i.e. GSM 900, 1800 and 2100 bands. The occupied spectrum calculation is done by setting a threshold level (a solid black line marked as a threshold level in all figures) as shown in Fig. 3 to Fig. 30. The half-power threshold is indicated by adding 3dBm in the minimum received power signal. If the measured APSD of a specific frequency is beyond the threshold level, then the certain frequency is called as occupied frequen, cy. The percentage of spectrum occupancy can be calculated as described in Eq. 1.

$$SO = \frac{P_T}{P} \times 100\% \qquad (1)$$

Where P_T denotes the number of frequency measurements that exceed the selected threshold level and P is the total number of frequency measurement points in the specific spectrum band.

A. Mobilink

Mobilink started its cellular services as a pioneer operator in Pakistan in 1994 [18]. Recently, they have won the 10 MHz spectrum in the 1800 MHz band. Mobilink and Warid merged into a single company on November 26, 2015. Also, the recent merger of Jazz and Warid increased their subscribers of about 49 million around the country.

The results of indoor and outdoor GSM 900 uplink and downlink spectrum occupancies are represented in Fig. 3, 4, 5 and 6, respectively. The measured results of indoor spectrum occupancy for both uplink and downlink spectrum bands are 61.794% and 61.4618%, respectively. These results conclude that a 5.82 MHz spectrum band is unutilized out of the total 15.2 MHz spectrum band. Similarly, the measured results of outdoor spectrum occupancy for both uplink and downlink spectrum bands are 58.1362% and 68.7708% respectively. These results conclude that a 5.47 MHz spectrum band is unutilized out of the total 15.2 MHz spectrum band.

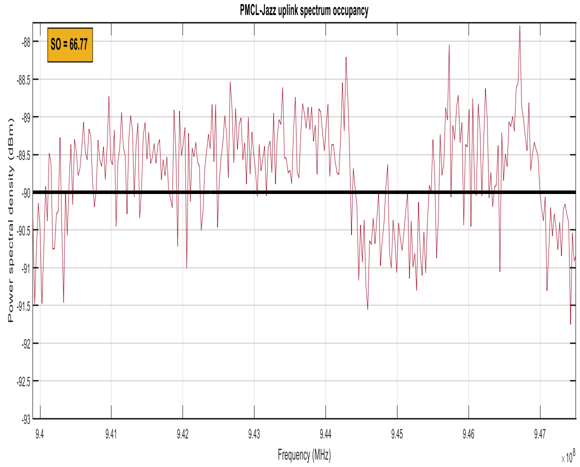


Fig 3. PMCL-Jazz uplink Spectrum Occupancy.

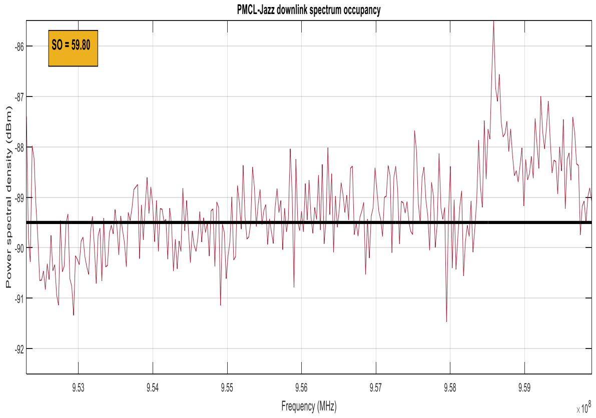


Fig 4. PMCL-Jazz Downlink Spectrum Occupancy.

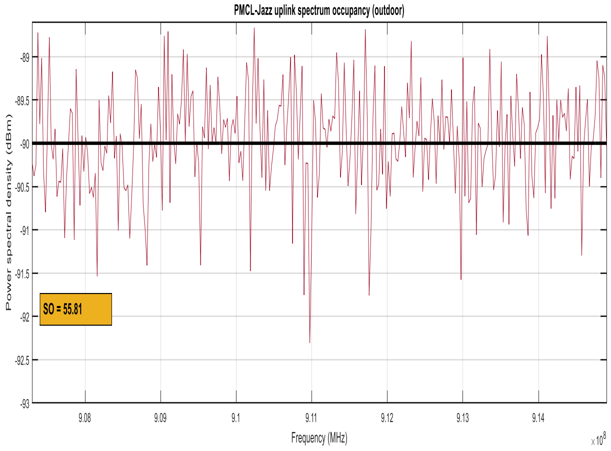


Fig 5. PMCL-Jazz uplink Spectrum Occupancy (Outdoor).

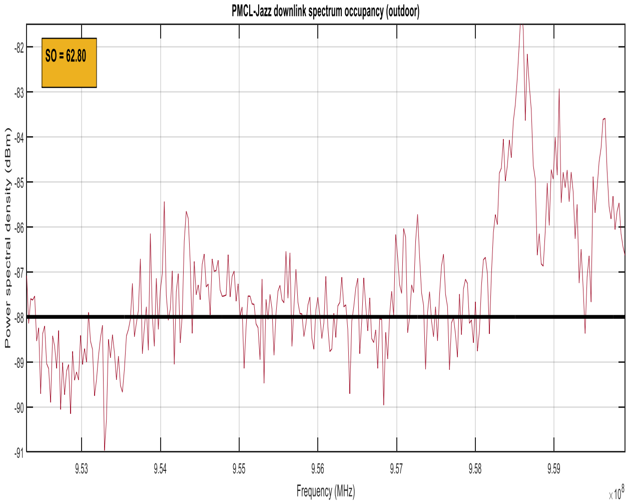


Fig 6. PMCL-Jazz Downlink Spectrum Occupancy (Outdoor).

B. UFONE

UFONE initiated its cellular services in Pakistan on January 29, 2001. Recently, it has launched its 4G services in major cities of Pakistan on February 9, 2019[19].

The measured results of indoor and outdoor GSM 900 uplink and downlink spectrum occupancies are illustrated in Fig. 7, 8, 9 and 10, respectively. The measured results of indoor spectrum occupancy for both uplink and downlink spectrum bands are 42.1927% and 91.0299% respectively. These results conclude that a 4.8 MHz spectrum band is unutilized out of the total 15.2 MHz spectrum band. Similarly, the measured results of outdoor spectrum occupancy for both uplink and downlink spectrum bands are 63.1229% and 86.711% respectively. These results conclude that a 3.8 MHz spectrum band is unutilized out of the total 15.2 MHz spectrum band.

C. Telenor

Telenor [17] is a multinational telecommunications company and it launched its GSM cellular services in the main cities of Pakistan on March 15, 2005. They have also won a 4G license spectrum of 10MHz block in 850 MHz in Pakistan.

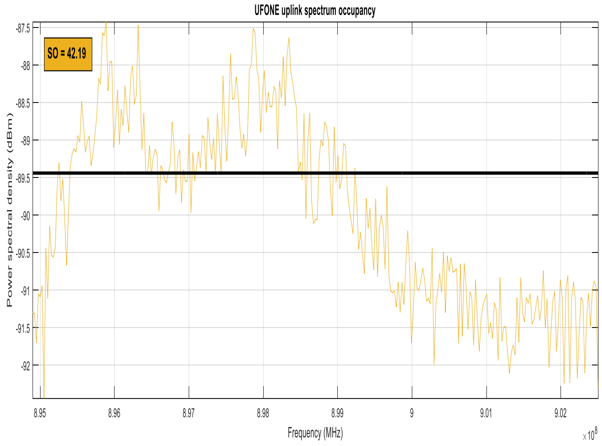


Fig 7. Ufone uplink Spectrum Occupancy.

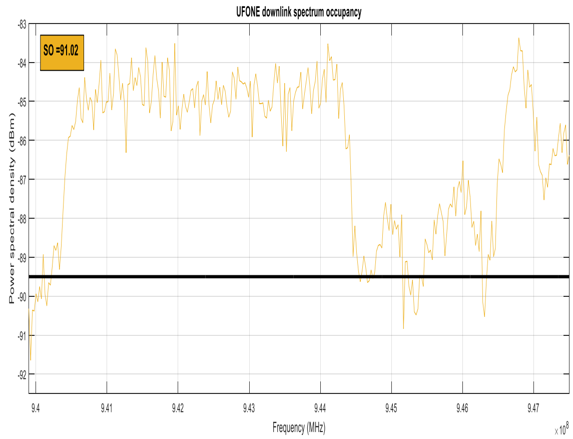


Fig 8. Ufone Downlink Spectrum Occupancy.

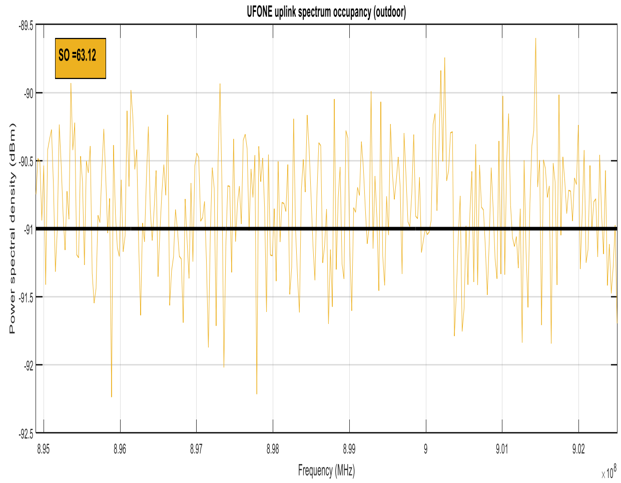


Fig 9. Ufone uplink Spectrum Occupancy (Outdoor).

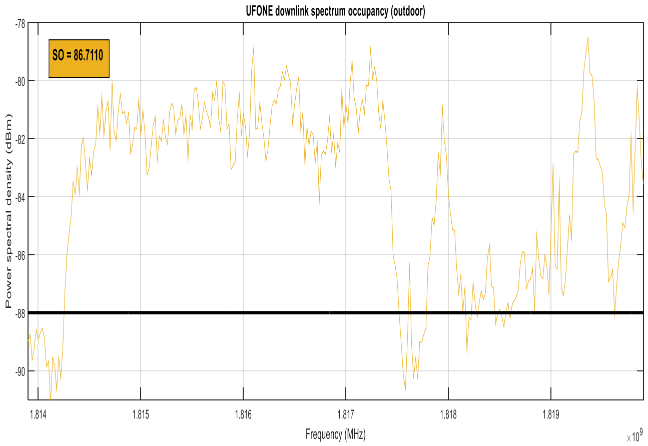


Fig 10. Ufone Downlink Spectrum Occupancy (Outdoor).

The results of indoor and outdoor GSM 900 uplink and downlink spectrum occupancies are represented in Fig. 11, 12, 13 and 14, respectively. The measured results of indoor spectrum occupancy for both uplink and downlink spectrum bands are 66.113% and 60.7973%, respectively. These results conclude that a 3.51 MHz spectrum band is unutilized out of the total 9.6 MHz spectrum band. Similarly, the measured results of outdoor spectrum occupancy for both uplink and downlink spectrum bands are 56.4784% and 66.7774% respectively. These results conclude that a 3.67 MHz spectrum band is unutilized out of the total 9.6 MHz spectrum band.

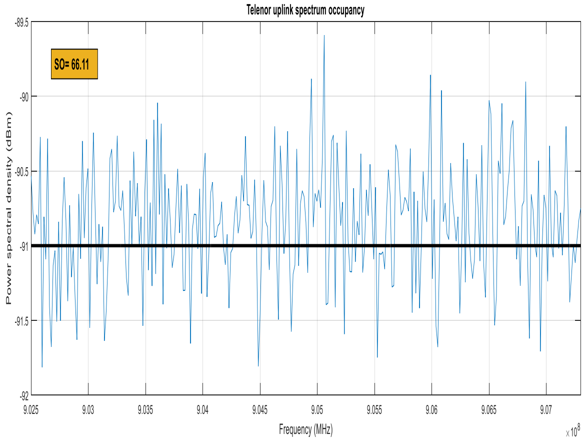


Fig 11. Telenor uplink Spectrum Occupancy.

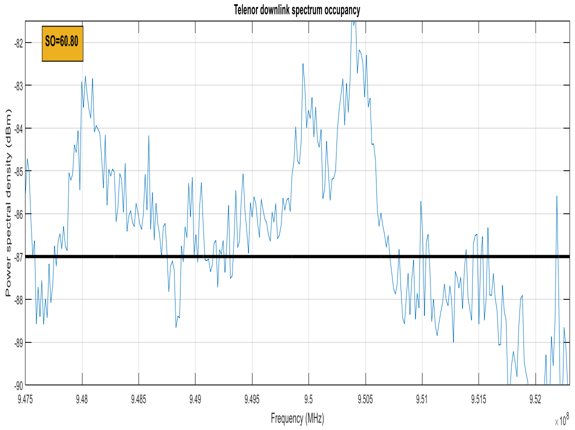


Fig 12. Telenor Downlink Spectrum Occupancy.

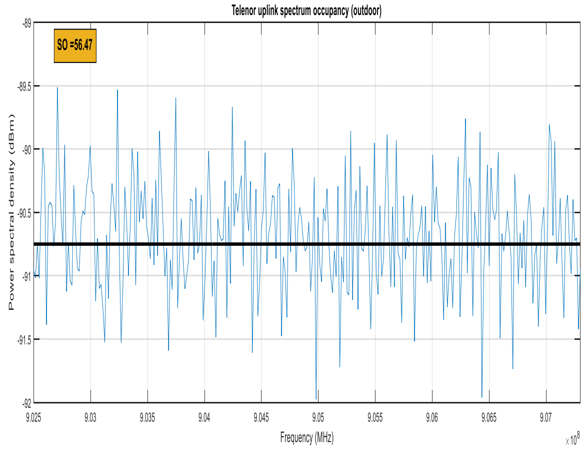


Fig 13. Telenor uplink Spectrum Occupancy (Outdoor).

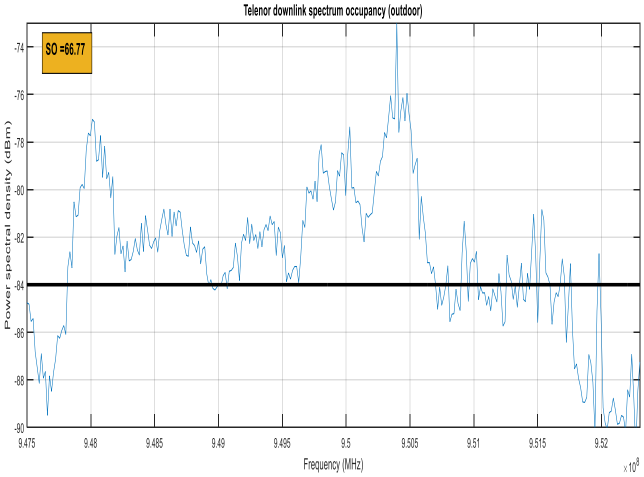


Fig 14. Telenor Downlink Spectrum Occupancy (Outdoor).

D. ZONG

Zong is the only cellular operator that is being owned by China and it launched its operation in Pakistan 1991 as Paktel by Cable & Wireless. Also, it is one of the first company which is allowed a free license to provide cellular phone services. Recently PTA has allotted a six months 5G trial license to check its coverage on non-commercial-basis [20].

The results of indoor and outdoor GSM 900 uplink and downlink spectrum occupancies are represented in Fig. 15, 16, 17 and 18, respectively. The measured results of indoor spectrum occupancy for both uplink and downlink spectrum bands are 69.103% and 88.0399% respectively. These results conclude that the 3.34 MHz spectrum band is unutilized out of the total 15.2 MHz spectrum band. Similarly, the measured results of outdoor spectrum occupancy for both uplink and downlink spectrum bands are 62.4585% and 81.3953% respectively. These results conclude that the 4.26 MHz spectrum band is unutilized out of the total 15.6 MHz spectrum band.

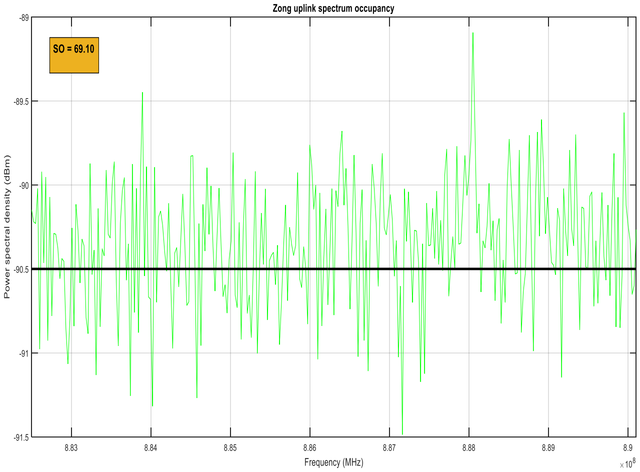


Fig 15. Zong uplink Spectrum Occupancy.

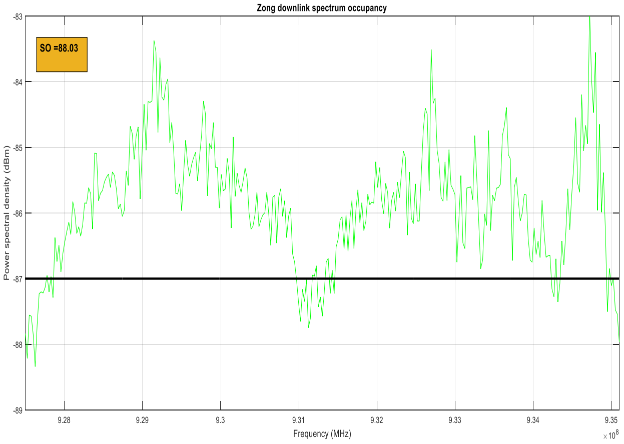


Fig 16. Zong Downlink Spectrum Occupancy.

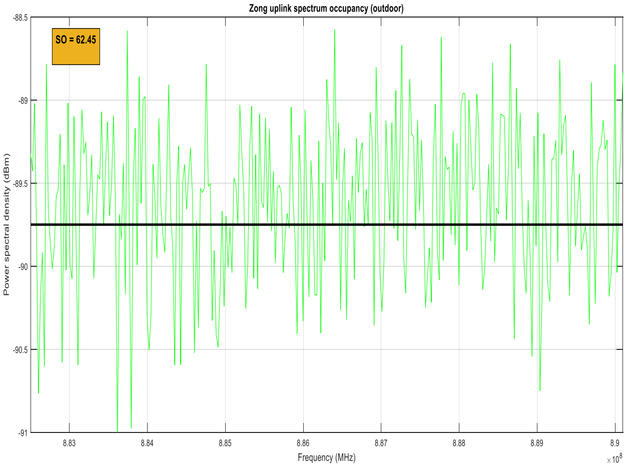


Fig 17. Zong uplink Spectrum Occupancy (Outdoor).

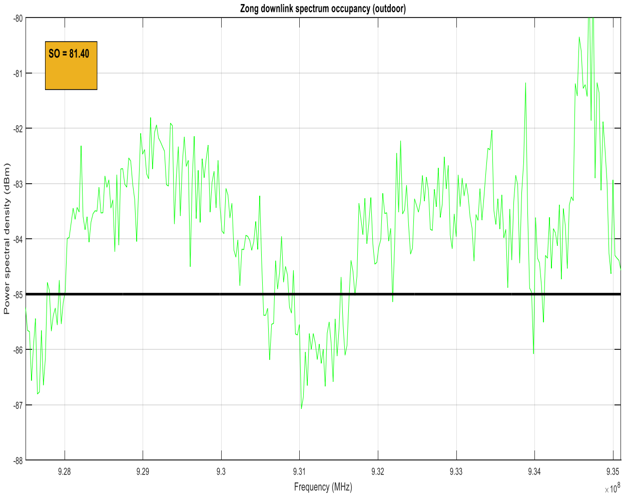


Fig 18. Zong Downlink Spectrum Occupancy (Outdoor).

E. GSM 900

The results of indoor and outdoor GSM 900 uplink and downlink spectrum occupancies are represented in Fig. 19, 20, 21 and 22, respectively. The measured results of indoor spectrum occupancy for both uplink and downlink spectrum bands are 57.1429% and 53.8206% respectively. These results conclude that 22.25 MHz spectrum band is unutilized out of the total 64.8 MHz spectrum band. Similarly, the measured results of outdoor spectrum occupancy for both uplink and downlink spectrum bands are 77.7409% and 56.8106% respectively. These results conclude that 16.35 MHz spectrum band is unutilized out of the total 64.8 MHz spectrum band.

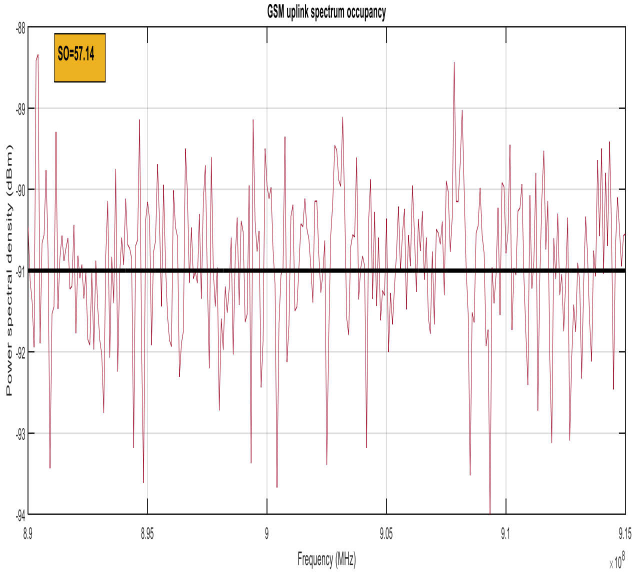


Fig 19. GSM 900 uplink Spectrum Occupancy.

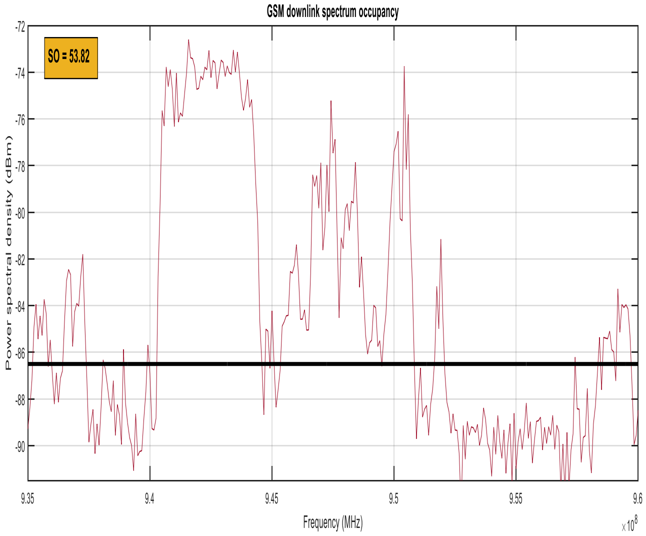


Fig 20. GSM 900 Downlink Spectrum Occupancy.

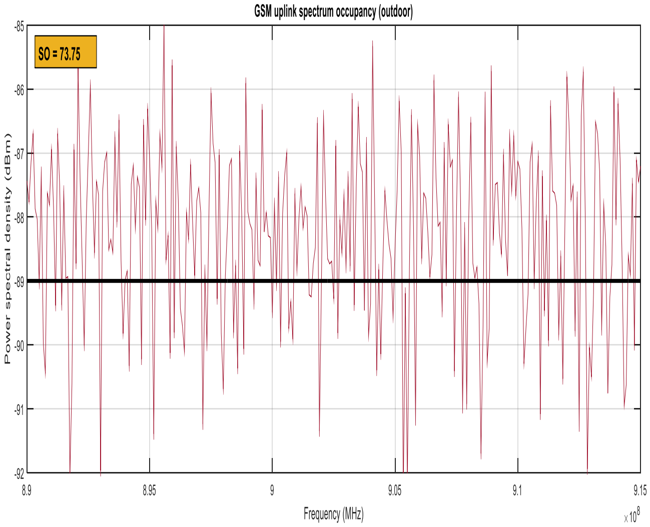


Fig 21. GSM 900 uplink Spectrum Occupancy (Outdoor).

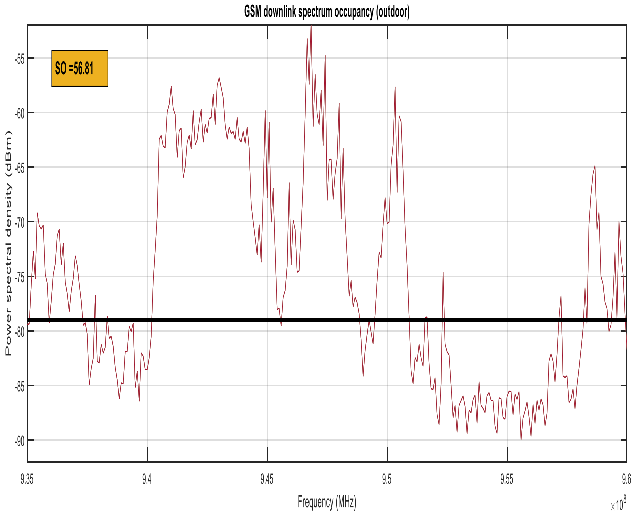


Fig 22. GSM 900 Downlink Spectrum Occupancy (Outdoor).

F. DCS 1800

The measured results of indoor and outdoor GSM 1800 uplink and downlink spectrum occupancies are represented in Fig. 23, 24, 25 and 26 respectively. The measured results of indoor spectrum occupancy for both uplink and downlink spectrum bands are 43.12% and 62.13% respectively. These results conclude that 58.93 MHz spectrum band is unutilized out of the total MHz spectrum band. Similarly, the measured results of outdoor spectrum occupancy for both uplink and downlink spectrum bands are 59.80% and 64.45% respectively. These results conclude that the 47.11 MHz spectrum band is unutilized out of the total 62.2 MHz spectrum band.

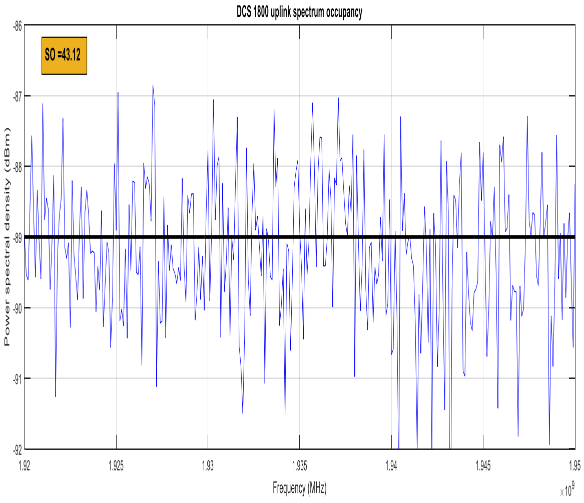


Fig 23. DCS 1800 uplink Spectrum Occupancy.

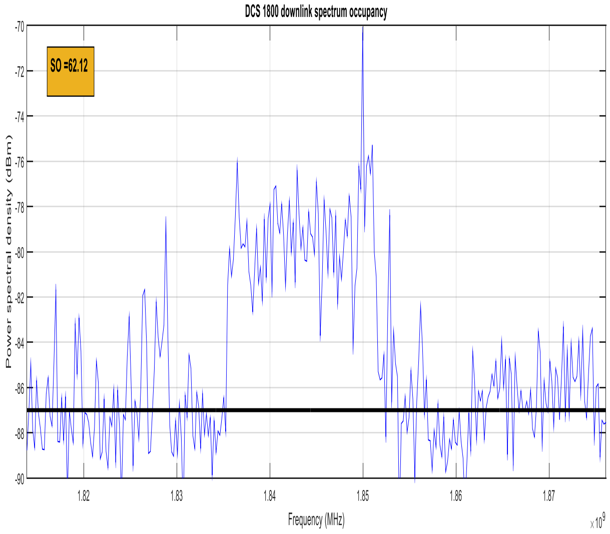


Fig 24. DCS 1800 Downlink Spectrum Occupancy.

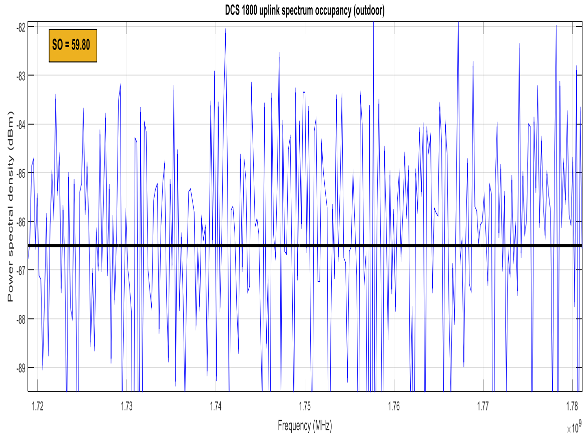


Fig 25. DCS 1800 uplink Spectrum Occupancy (Outdoor).

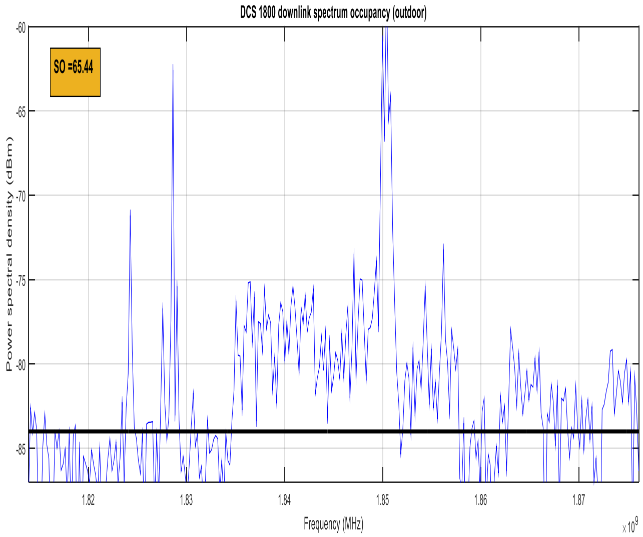


Fig 26. DCS 1800 Downlink Spectrum Occupancy (Outdoor).

G. 2100 Band

The results of indoor and outdoor 2100 uplink and downlink spectrum occupancies are represented in Fig. 27, 28, 29 and 30, respectively. The measured results of indoor spectrum occupancy for both uplink and downlink spectrum bands are 43.52% and 77.07%, respectively. These results conclude that the 23.81 MHz spectrum band is unutilized out of the total 30 MHz spectrum band. Similarly, the measured results of outdoor spectrum occupancy for both uplink and downlink spectrum bands are 59.8% and 86.71% respectively. These results conclude that the 16.05 MHz spectrum band is unutilized out of the total 30 MHz spectrum band.

Spectrum occupancy and their related unused spectrum of all the GSM cellular operators and GSM 900, 1800 and 2100 bands are illustrated in Table III to Table IX.

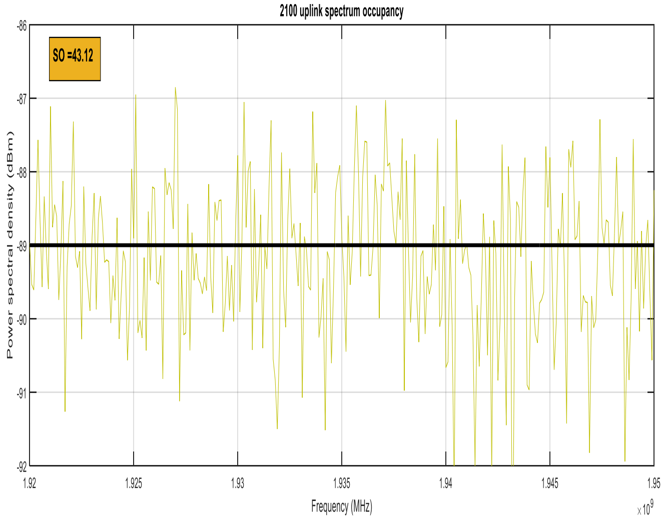


Fig 27. 2100 uplink Spectrum Occupancy.

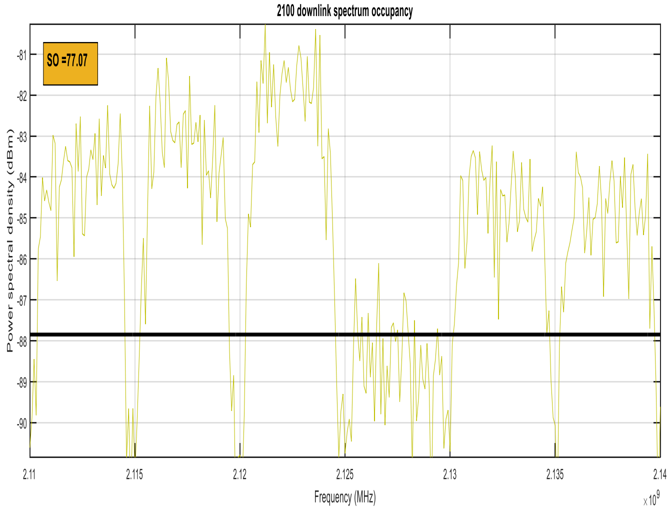


Fig 28. 2100 Downlink Spectrum Occupancy.

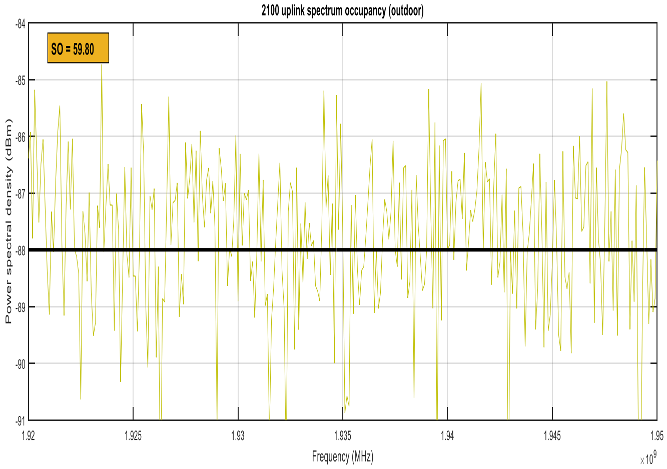


Fig 29. 2100 uplink Spectrum Occupancy (Outdoor).

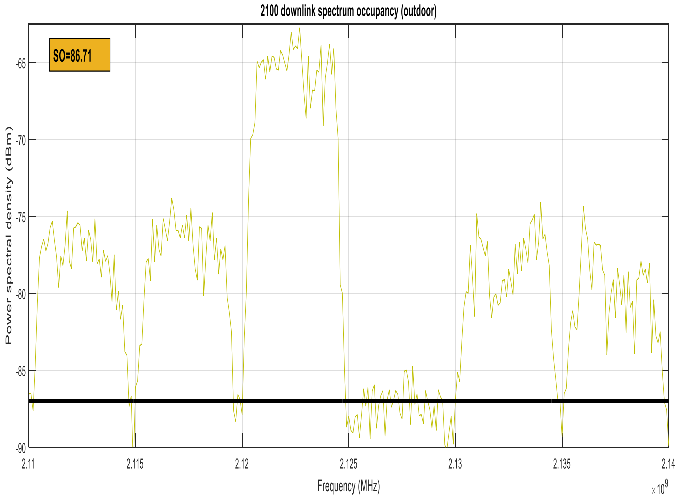


Fig 30. 2100 Downlink Spectrum Occupancy (Outdoor).

TABLE III. A SUMMARY OF SPECTRUM OCCUPANCY AND UNUSED SPECTRUM (INDOOR: GSM900)

GSM SERVICE PROVIDERS		Spectrum occupancy (%)	Unused Spectrum (MHz)
Jazz (Mobilink)	Uplink	61.794	2.90
	Downlink	61.4618	2.92
Ufone	Uplink	42.1927	4.30
	Downlink	91.0299	0.68
Telenor	Uplink	66.113	1.63
	Downlink	60.7973	1.88
Jazz (Warid)	Uplink	69.7674	1.45
	Downlink	73.0897	1.29
Zong	Uplink	69.103	2.34
	Downlink	88.0399	1.00
GSM 900	Uplink	57.1429	10.71
	Downlink	53.8206	11.54

TABLE IV. A SUMMARY OF SPECTRUM OCCUPANCY AND UNUSED SPECTRUM (OUTDOOR: GSM900)

GSM SERVICE PROVIDERS		Spectrum occupancy (%)	Unused Spectrum (MHz)
Jazz (Mobilink)	Uplink	59.1362	3.1
	Downlink	68.7708	2.37
Ufone	Uplink	63.1229	2.8
	Downlink	86.711	1
Telenor	Uplink	56.4784	2.08
	Downlink	66.7774	1.59
Jazz (Warid)	Uplink	68.4385	1.51
	Downlink	63.1229	1.77
Zong	Uplink	62.4585	2.85
	Downlink	81.3953	1.41
GSM 900	Uplink	77.7409	5.56
	Downlink	56.8106	10.79

TABLE V. THE SA SETTINGS FOR GSM SERVICE PROVIDERS DCS 1800 BAND IN PAKISTAN

SA PARAMETERS		Frequency range (MHz)	Center frequency (MHz)	Frequency span (MHz)	Resolution Bandwidth (KHz)	Video Bandwidth (KHz)	Sweep time (ms)	Number of frequency measurement points
DCS 1800	Uplink	1718.9-1781.1	1750	62.2	200	200	250	301
	Downlink	1813.9-1876.1	1845	62.2	200	200	250	301

TABLE VI. DCS 1800 INDOOR AND OUTDOOR

A SUMMARY OF SPECTRUM OCCUPANCY AND UNUSED SPECTRUM (INDOOR: DCS1800)				A SUMMARY OF SPECTRUM OCCUPANCY AND UNUSED SPECTRUM (OUTDOOR: DCS1800)			
GSM Service Providers		Spectrum occupancy (%)	Unused Spectrum (MHz)	GSM Service Providers		Spectrum occupancy (%)	Unused Spectrum (MHz)
DCS 1800	Uplink	43.12	35.37	DCS 1800	Uplink	59.8	25
	Downlink	62.13	23.56		Downlink	64.45	22.11

TABLE VII. SPECTRUM ANALYZER SETTINGS 2100 INDOOR AND OUTDOOR

THE SA SETTINGS FOR GSM SERVICE PROVIDERS 2100 BANDS IN PAKISTAN								
SA parameters		Frequency range (MHz)	Center frequency (MHz)	Frequency span (MHz)	Resolution Bandwidth (KHz)	Video Bandwidth (KHz)	Sweep time (ms)	Number of frequency measurement points
GSM Service Providers								
2100	Uplink	1920-1950	1935	30	200	200	250	301
	Downlink	2110-2140	2125	30	200	200	250	301

TABLE VIII. SUMMARY OF 2100 INDOOR AND OUTDOOR

A SUMMARY OF SPECTRUM OCCUPANCY AND UNUSED SPECTRUM (INDOOR)			
GSM Service Providers		Spectrum occupancy (%)	Unused Spectrum (MHz)
2100	Uplink	43.52	16.94
	Downlink	77.07	6.87

TABLE IX. SUMMARY OF 2100 OUTDOOR

A SUMMARY OF SPECTRUM OCCUPANCY AND UNUSED SPECTRUM (OUTDOOR)			
GSM Service Providers		Spectrum occupancy (%)	Unused Spectrum (MHz)
2100	Uplink	59.8	12.06
	Downlink	86.71	3.99

H. Smart Network Sharing

Undoubtedly, the prompt advancement in Telecommunication compels every Mobile operator in Pakistan to be more innovative for the growth of the business. Secondly, cellular customers in Pakistan are increasing at a very fast pace, especially data users. Incumbent operators, Jazz, Telenor, Zong, and Ufone are doing their best efforts to provide not only better services to existing customers but also in a bid to improve their coverage in every corner of the country. Nevertheless, all the mobile service providers are facing the difficult task of significantly increasing the capacity to meet projected demand while keeping CAPEX and OPEX down. In order to achieve the same, infrastructure sharing become a key consideration in operators' planning of the evolution of their networks in the current era. Fig. 31 is showing the number of mobile subscribers in Pakistan until the last quarter (November-2019).

Additionally, to further clarify this, the subscribers of every mobile operator are shown in Fig. 32. These pictures clearly depict the growth in subscribers with each year passing on.

A survey conducted in this project to find out the insights and status of network sharing in Pakistan among operators. Mostly, mobile operators are doing Passive infrastructure sharing due to government regulations. There are four models of Passive infrastructure sharing; model: 1) (Only site space and tower is shared), model: 2) (Site space, tower, and Commercial power is shared), model: 3) (Site Space, tower, Commercial power, and Genset backup is shared), model: 4) (Site Space, tower, commercial power, Genset backup, and DC power is shared). With respect to market share, Jazz is leading in the market and stands number one operator with 37%, while Telenor on second in a row with 28%, whereas Zong is on the third number with 19% and Ufone is on the fourth position with 15%. With these business opportunities and challenges, mobile operators are witnessing the hardest era in terms of cost. Therefore, to reduce these expenditures, network sharing proved to be the best solution for mobile operators in Pakistan. As per findings, the number of base transceiver stations (BTS) of four operators in Pakistan are shown in below Table X.

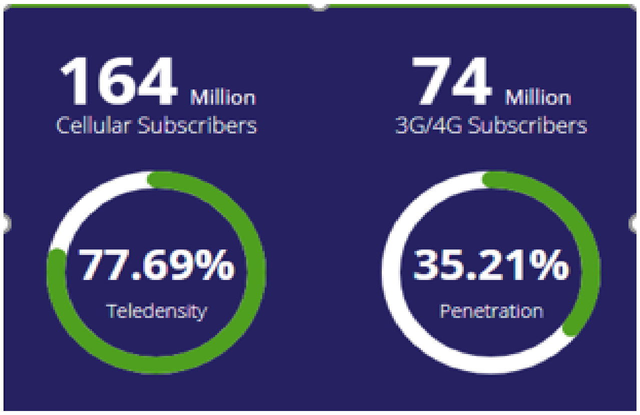


Fig 31. Number of Mobile subscriber in Pakistan.

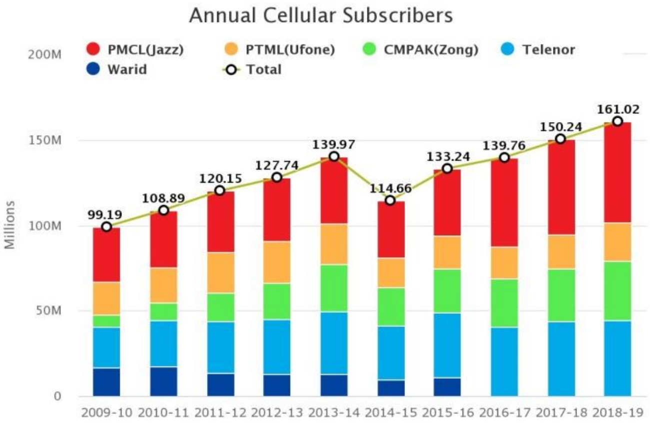


Fig 32. Annual trend of the cellular subscribers in Pakistan.

Interestingly, Telenor is leading among all operators when it comes to network sharing in Pakistan with 37.52% of Telenor network is shared. While Ufone is on the second number in this competition with 27.69% of its shared network. Zong is on third with 25.78% of shared network and Jazz is the last one with 22.63% network sharing. This clearly shows the incline of all operators towards smart network sharing in the Pakistan Telecom industry as illustrated in Table X and Fig. 33.

As one of the biggest advantages of the network, sharing is cost-saving, therefore, the details of cost savings provided here. As per this research, operators bear the expenditure of base transceiver station in Pakistan consists of annual rental payment, electricity billing, security expenditure, operation, and maintenance expenditure. Table XI illustrates these details of expenditure.

TABLE X. MONTHLY EXPENDITURE DETAILS IN 2019

Mobile Operator	**Total number of BTS**
Jazz	13700
Telenor	12199
Zong	12590
Ufone	9500

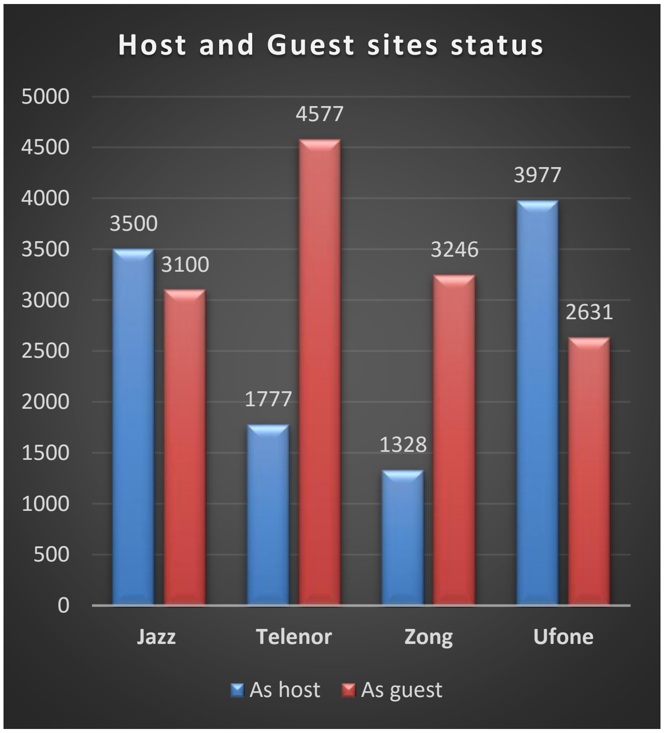


Fig 33. Status of sharing (Host & Guest) sites of Operators.

TABLE XI. MONTHLY EXPENDITURE DETAILS IN 2019

Monthly expenditure details in 2019		
Details	Without sharing/Independent site Expenditure details	With Sharing/co-location
Average Rent	20 to 50 thousand per month	0
Electricity Expenditure	70 to 80 thousand per month	0
Security Expenditure	12000	0
O&M of CP & DG expenditure	7,000	0
Total	109000 to 149000	0

IV. CONCLUSION

In this paper, we have carried out the spectrum occupancy measurements through SA and plot them in the MATLAB against frequency vs power spectral density. We have measured the spectrum occupancies of three bands i.e. 900, 1800 and 2100 bands for both indoor as well as outdoor spectrum bands and found out that the majority of the spectrum band is unutilized. The results of measured spectrum occupancies are illustrated in Table III to Table IX. In addition, it is noted that the overall bandwidth utilization ratio of all the cellular service providers is not 100%, which can be utilized with the help of CR. Thus, CR can be the optimal solution to improve the spectrum utilization ratio of all the three bands of uplink and downlink spectrum bands in Pakistan. Moreover, the Network sharing became the norm for Telecom industry to control cost and expenditure.

V. Future Work

This research work can give a promising result to open the doors of a new era of wireless communication in Pakistan in terms of network sharing. Spectrum bands will become technological free/neutral as per the demand of this day and spectrum occupancy will play a vital role for it. Furthermore, it is likely that 3G infrastructure may vanish in future, as the 3G architecture entails extensive hardware replacing when shifting from 2G to 3G (expensive). 4G/LTE is an all IP network and do not need conventional hardware though it only requires switches and servers. It is therefore likely that a third party vendor/contractor may provide site allocation and infrastructure to cellular operators, reducing OPEX and CAPEX by the concept of smart network sharing.

References

[1] H. Li, X. Ding, Y. Yang, Z. Xie, and G. Zhang, “Online Spectrum Prediction with Adaptive Threshold Quantization,” IEEE Access, vol. 7, pp. 174325–174334, 2019, DOI: 10.1109/ACCESS.2019.2957335.

[2] H. Qi, S. Member, X. Zhang, and S. Member, “Low-Complexity Subspace-Aided Compressive,” IEEE Trans. Veh. Technol., vol. 68, no. 12, pp. 11762–11777, 2019, DOI: 10.1109/TVT.2019.2937649.

[3] Y. Luo, J. Dang, and Z. Song, “Optimal Compressive Spectrum Sensing Based on Sparsity Order Estimation in Wideband Cognitive Radios,” IEEE Trans. Veh. Technol., vol. 68, no. 12, pp. 12094–12106, 2019, DOI: 10.1109/TVT.2019.2948966.

[4] Y.-H. Liu, S. Sheelavant, M. Mercuri, P. Mateman, and M. Babaie, “An Ultralow Power Burst-Chirp UWB Radar Transceiver for Indoor Vital Signs and Occupancy Sensing in 40-nm CMOS,” IEEE Solid-State Circuits Lett., vol. 2, no. 11, pp. 256–259, 2019, DOI: 10.1109/lssc.2019.2951423.

[5] S. Ali, Z. Chen, and F. Yin, “Spectrum Occupancy of Cellular Networks in Pakistan for Cognitive Radio — Measurements using Spectrum Analyzer,” Int. J. Inf. Electron. Eng., vol. 6, no. 1, pp. 26–31, 2015, DOI: 10.18178/ijiee.2016.6.1.588.

[6] D. Capriglione, G. Cerro, L. Ferrigno, and G. Miele, “Effects of Real Instrument on Performance of an Energy Detection-Based Spectrum Sensing Method,” vol. 68, no. 5, pp. 1302–1312, 2019.

[7] M. Ozturk, M. Akram, S. Hussain, and M. A. Imran, “Novel QoS-Aware Proactive Spectrum Access Techniques for Cognitive Radio Using Machine Learning,” IEEE Access, vol. 7, pp. 70811–70827, 2019, DOI: 10.1109/ACCESS.2019.2918380.

[8] N. A. El-Alfi, H. M. Abdel-Atty, and M. A. Mohamed, “Sub-Nyquist Cyclostationary Detection of GFDM for Wideband Spectrum Sensing,” IEEE Access, vol. 7, pp. 86403–86411, 2019, DOI: 10.1109/ACCESS.2019.2925047.

[9] D. V. Kalkitware, “Wireless Body area networks for patient monitoring,” Eng. Technol., vol. 2, no. 6, pp. 215–222, 2016, DOI: 10.1049/iet-com.

[10] M. V. Lipski and R. M. Narayanan, “Applying Periodic Retraining to Survival Analysis-Based Dynamic Spectrum Access Algorithms,” Proc. - IEEE Mil. Commun. Conf. MILCOM, vol. 2019-October, pp. 871–876, 2019, DOI: 10.1109/MILCOM.2018.8599767.

[11] G. Cerro and G. Miele, “A stand-alone sensor for spectrum occupancy monitoring in dynamic spectrum access framework,” SAS 2019 - 2019 IEEE Sensors Appl. Symp. Conf. Proc., pp. 1–6, 2019, DOI: 10.1109/SAS.2019.8706108.

[12] L. N. Man, S. Committee, and I. Computer, Part 22 : Cognitive Wireless RAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications : Policies and Procedures for Operation in the TV Bands IEEE Computer Society, vol. 2015, no. July. 2011.

[13] A. Bousia, E. Kartsakli, A. Antonopoulos, L. Alonso, and C. Verikoukis, “Sharing the small cells for energy-efficient networking: How much does it cost?,” 2014 IEEE Glob. Commun. Conf. GLOBECOM 2014, pp. 2649–2654, 2014, DOI: 10.1109/GLOCOM.2014.7037207.

[14] T. International et al., “Mobile Cellular Network Infrastructure Sharing Models Among Gsm,” vol. 3, no. 4, pp. 12–13, 2015.

[15] L. Anchora et al., “Resource allocation and management in multi-operator cellular networks with shared physical resources,” Proc. Int. Symp. Wirel. Commun. Syst., pp. 296–300, 2012, DOI: 10.1109/ISWCS.2012.6328377.

[16] L. Chihana and D. Banda, “Telecommunication Tower Sharing Effects on Network Providers in Zambia,” vol. 7, no. 4, pp. 89–93, 2017, DOI: 10.5923/j.scit.20170704.01.

[17] “Breaking: Telenor Wins 4G Spectrum in Pakistan,” 2016. [Online]. Available: https://propakistani.pk/2016/06/09/breaking-telenor-wins-4g-spectrum-in-pakistan/. [Accessed: 10-Feb-2020].

[18] “Mobilink, Warid become Jazz after the merger,” 2017. [Online]. Available: https://www.dawn.com/news/1313181. [Accessed: 17-Feb-2020].

[19] “Ufone is Going 4G/LTE in Pakistan,” 2019. [Online]. Available: https://propakistani.pk/2019/02/09/ufone-is-going-4g-lte-in-pakistan/. [Accessed: 17-Feb-2020].

[20] “Pakistan grants Jazz, Zong 5G trial licenses,” 2020. [Online]. Available: https://www.mobileworldlive.com/asia/asia-news/pakistan-grants-jazz-zong-5g-trial-licences/. [Accessed: 17-Feb-2020].

Analysis on the Requirements of Computational Thinking Skills to Overcome the Difficulties in Learning Programming

Karimah Mohd Yusoff[1], Noraidah Sahari Ashaari[2], Tengku Siti Meriam Tengku Wook[3], Noorazeen Mohd Ali[4]
Matriculation Division, Ministry of Education Malaysia, Putrajaya, Malaysia[1]
Software Technology and Management System, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia[2, 3, 4]

***Abstract*—Programming has evolved as an effort to strengthen science, technology, engineering and mathematics (STEM). Programming is a complex process, especially for novices, since it requires problem-solving skills to solve problems of developing algorithms and programme codes. Problem-solving competencies, which are necessary as 21st-century skills, include a set of cognitive skills that are related to problem-solving and programme development or specifically known as computational thinking (CT) skills. In particular, this study quantitatively assessed the computational thinking skills in the context of programming, specifically on the difficulties in learning programming. From the perspectives of the instructors, the survey results highlighted the need to implement CT skills as an approach in teaching and learning programming. A model for teaching and learning programming is necessary as a guide for instructors in the teaching and learning process of programming.**

Keywords—Problem-solving; STEM; difficulties in learning programming; cognitive; novice

I. Introduction

Job opportunities and daily activities that involve computers have encouraged students to pursue computing, such as computer engineering, computer science, information science, and software engineering, as a career [1]. The U.S. Bureau of Labour Statistics reported that computing represented 71% of the careers in science, technology, engineering and mathematics (STEM) by 2018 [2]. The programming curriculum has gained growing attention given the significance of computing in meeting the current needs and STEM agenda. Programming is emphasized in schools and even at the pre-university level in order to provide students with a good knowledge base and programming skills. However, it is a challenging and complex process to learn programming [3][4][1], since it requires good cognitive ability. Novice programmers often face problems during the introductory course of programming [5][6] that may cause hesitation to pursue the advanced courses of programming. This scenario shows that the early mastery of programming serves as a catalyst for students to consider courses related to programming at a later stage.

In general, programming is viewed as a means of producing computer programmes. Basically, programming solves real-world problems through computer programmes. The implementation of the identified solution involves several steps, which are as follows: (1) formulate a problem; (2) design a solution by generating an algorithm; (3) translate the algorithm into a programme code; (4) test and evaluate the complete programme. Although studies have introduced several programming methods and approaches to assist novice programmers, not all focus on the programming steps involved in solving problems. Furthermore, most of the past studies focused on mastering the concepts of programming, such as learning using MicroWorlds, game-based learning, story-based learning, and visualisation tools.

For instance, the use of MicroWorlds, such as Alice, Greenfoot, Marine Biology Case Study, Scratch, and Turtle Graphics, provides a user-based interface (GUI) that introduces basic programming concepts to novices. Although such approach can build problem-solving skills, it mainly focuses on implementing solutions in specific programming languages. In addition, it does not develop one’s ability to formulate problems, design solutions, and generate algorithms that are often necessary for large, complex problems or involve multiple processes. Besides, games-based learning approaches are considered to provide students with a lot of fun while learning, but it also focusses on programming concepts [7]. Meanwhile, work-based learning approaches involve problem solving in programming that can help to reduce the learners’ cognitive load by performing solutions based on work examples [8][9][10]. Most importantly, the learning process must involve learners themselves, where the learning process is designed according to their competencies. Learners are also required to engage in a group discussion to discuss, interact and provide feedback, and guide their peers.

Accordingly, the idea of computational thinking (CT) already existed during the early 1950s and has gained growing attention of educators and researchers over the past decade. The term “computational thinking” was first used by Seymour Papert in 1980 and 1996 [11][12]. Following that, Wing [13] formally introduced CT as an approach to solving problems, designing systems, and understanding human behaviour, which reflects the basic concepts of computing. The discussion of CT in the literature is often associated with problem-solving [14].

CT is an important and necessary way of thinking for computer programmers and other professionals in STEM [15]. CT skills, which include decomposition, abstraction, pattern-recognition, algorithm, logical reasoning, and assessment (or evaluation) skills, are cognitive skills that can be used in the teaching and learning process of programming that typically involves problem-solving. The computational skills of CT derived from computer science have the potential to be used for problem solving in all disciplines [16]. Through these computational skills, one can be better at solving problems and can identify problems and apply a smart approach to solve the problems [17]. Problem solving in programming involves several processes that can be implemented using appropriate CT skills. In general, there are two main phases of problem-solving in programming. The first phase is to generate an algorithm whereas the second phase is to develop a programme. Both phases coincide with the role of CT as a thought process that involves formulating problems and solutions in a form that can be effectively implemented by the information processing agents [18] such as computers.

Although CT is an ideal teaching and learning approach that can help with the curriculum problems [19], its implementation, to date, focuses on K-12 only [20][21][22][23]. Studies have revealed limited CT implementation for higher learning, particularly at the pre-university level. Practical research on teaching CT skills at the higher education level has been continuously implemented in computer science and STEM [17]. The difficulties in learning programming have been a topic of discussion among educators and researchers.

The difficulties in learning programming were widely explored in past studies [24][25][26][27][28][29][30][31]. Besides that, Du Boulay [3], Robins, Rountree, and Rountree [4], and Qian and Lehman [32] also reviewed the difficulties in learning programming. Some of the identified difficulties included designing solution plans, developing algorithms, syntactic mastery, writing and evaluating programmes, cognitive requirements, and limited programming ability. The problem-solving approach can be implemented using CT skills according to the required role. For instance, abstraction skills play a role in identifying and retrieving relevant information to determine key ideas and reduce unnecessary information. Besides that, decomposition skills help to decompose complex problems (that involve several processes) according to the process, which makes problem solving easier, as the problems can now be solved in parts. Meanwhile, through pattern-recognition skills, programmers can observe the patterns, trends, and regularity of data by observing the similarities and differences with other problems. There are also the algorithm skills that involve a set of rules and instructions to execute tasks or address any problem-solving needs in programming. These skills can help programmers to develop computer algorithms, specifically the step-by-step solutions into forms that can be implemented by a computer. Apart from that, there are logical reasoning skills by analysing and studying facts based on accurate and clear-thinking approach. After all, problem-solving involves logic. Last but not least, the needs of each solution should be evaluated, which highlights the important role of evaluation (or assessment) skills in determining the adequacy of an algorithm, system, or process in serving its purpose of meeting the needs.

In short, CT plays an important role in learning and solving problems and computerizes thinking in all disciplines [33] given its significance as the core of STEM for solving problems and designing large, complex systems [12]. Focusing on the significance of 21st-century skills, learners need to master CT skills in order to solve problems based on the principles of computer science [14]. Clearly, CT is a thinking approach to develop problem-solving skills using the basic concepts of computing. In view of the above, the current study focused on high-level approaches to solve problems in programming using CT proficiency.

Due to the nature of programming that closely related with the problem-solving, computational thinking skills are potential to overcome the difficulties in programming. This study reviewed related literature on the difficulties in learning programming, how these difficulties are linked to computational thinking, and the need for computational thinking in learning programming. The obtained findings of this study were expected to benefit both instructors and students or novice programmers, especially in the preparation of an effective teaching and learning approach to programming. Hence, in this paper, author concerns to study the needs of computational thinking skills to overcome the difficulties in learning programming.

II. Research Purpose

For more than a decade, the use of CT as an approach to problem-based learning has prompted researchers to explore its use in computer science, especially for programming [34][35][36][37]. In addition, there is an increasing need to understand the role and skills of CT and identify the need to use CT skills in problem-solving and programme development. With that, this study aimed to identify the difficulties in learning programming and the required CT skills among learners in order to facilitate the teaching and learning process of programming. In particular, this study addressed the following research questions:

1) What are the common difficulties in learning programming among students?

2) What are the CT skills that are associated with the identified difficulties in learning programming?

3) Do the instructors face the identified difficulties?

4) What are the CT skills needed to overcome the difficulties in learning programming?

III. Methodology

A. *Mapping the Difficulties in Learning Programming with Computational Thinking Skills*

With respect to the purpose of this study, the difficulties in learning programming, especially among novice programmers, were reviewed. For this study, the difficulties in learning programming were first mapped based on the review of key literature, analysis of documents, and the elements of difficulties in relation to the CT skills. Fig. 1 illustrates the applied mapping method.

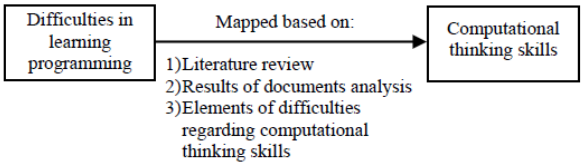

Difficulties in learning programming
Mapped based on:
Computational thinking skills
1)Literature review
2)Results of documents analysis
3)Elements of difficulties regarding computational thinking skills

Fig. 1. Mapping Method for Difficulties in Learning Programming with Computational Thinking Skills.

Meanwhile, Table I presents the mapping results, which showed that the difficulties in learning programming are entirely related to all CT skills, such as abstraction, decomposition, pattern-recognition, algorithms, logical reasoning, and assessment (or evaluation) skills. These initial findings demonstrated the significance of CT skills in developing the required instrument for the ensuing survey for this study.

TABLE I. RELATIONSHIP OF DIFFICULTIES IN LEARNING PROGRAMMING WITH COMPUTATIONAL THINKING SKILLS

References	***Difficulties in programming***	***Justifications of mapping to computational thinking (CT) skills***	***Computational Thinking (CT) Skills***					
			Decomposition	***Abstraction***	***Pattern-Recognition***	***Algorithm***	***Logical Reasoning***	***Assessment or Evaluation***
Qian and Lehman [32]	1) Not familiar with the syntax	1) Syntax is closely related to the closeness of mapping, which is a relationship between the programming languages and students' existing knowledge of the concepts used. The existing knowledge refers to knowledge in programming or other knowledge that applies the same concepts to programming. It is related to the pattern-recognition skills. 2) The existing knowledge used to predict is part of logical reasoning [38]. 3) Evaluation skills are indirectly required when logical reasoning is used to predict.			√		√	√
	2) Lack of ability in mathematics	1) The ability in mathematics is related to cognitive skills. 2) CT skills are a set of cognitive skills. These skills are in tandem with problem-solving skills for mathematics. 3) Problem-solving for mathematics requires strategy; perform solution sequentially; and involve logical reasoning and evaluation skills.	√	√	√	√	√	√
	3) Lack of mental model to implement codes	1) This is related to the concepts of programming. It involves cognitive skills to master it. 2) The use of analogies in daily life can help to understand the implementation of codes, as variables can hold one value at a time; each statement ends with a semicolon and each repeated process has a numerator to track it. This is related to pattern-recognition, logical reasoning, and evaluation skills. 3) Students in the study were found to face problem to understand how codes work. This issue is related to algorithm since logic in programming is in line with the logic flow of algorithm.			√	√	√	√
	4) Lack of strategic knowledge	1) This leads to difficulties in designing solutions, writing programmes, and resolving errors. 2) These issues are related to decomposition, abstraction, pattern-recognition, algorithm, logical reasoning, and evaluation skills.	√	√	√	√	√	√
	5) Lack of existing knowledge in programming	1) The existing knowledge is related to pattern-recognition, logical thinking, and assessment skills. 2) Existing knowledge that is used to predict is part of logical reasoning [38].			√		√	√

References	***Difficulties in programming***	***Justifications of mapping to computational thinking (CT) skills***	***Computational Thinking (CT) Skills***					
			Decomposition	***Abstraction***	***Pattern-Recognition***	***Algorithm***	***Logical Reasoning***	***Assessment or Evaluation***
	6) Difficult to solve complex tasks	1) Complex tasks involve students' cognitive load. 2) It is related to strategies of problem solving with respect to decomposition skill to decompose complex problems, abstraction skill to simplify the problem by removing unimportant information while not losing important information, and the ability to adapt other methods of almost similar solutions. 3) Students in the study were found to demonstrate the tendency of making mistakes when it comes to managing complex tasks. 4) Involves logical thinking and assessment skills.	√	√	√	√	√	√
Kwon [31]	7) Difficulty in designing a solution plan	1) Some of the proposed strategies are: a) Decompose the complex task in parts b) Retrieve relevant information to perform solution c) Identify other similar problem and adapt its solution 2) Abstraction and algorithm are closely related to the capability of problem solving [39]. 3) Students in the study were found to fully understand the problem and have the ability to describe the solution but difficult to figure out a solution or develop instructions that can be implemented by a computer. Related to the ability to generate or develop algorithms. Logical thinking skill is also involved in developing algorithms.	√	√	√	√	√	
Papadopoulos and Tegos [29]	8) Lack of problem solving	1) Problem solving involves planning such in problem (7) to perform solution in the form of algorithm.	√	√	√	√	√	
	9) Lack of CT skills	2) Indirectly related to CT skills	√	√	√	√	√	√
Siti Rosminah and Ahmad Zamzuri [30]	10) Difficult to understand the basic concepts of programming structures and programme design	1) Based on the literature, this is related to the existing knowledge. This problem can be solved using the phenomenon in daily life as a comparison to understand the concepts of programming structure. It is related to pattern recognition skill. 2) Designing programmes is related to the ability to develop an algorithm or symbolic languages that describe solutions in programme codes. 3) Logical reasoning is used to predict the aftermath of recognising the patterns and generate the algorithms. Logical reasoning assesses whether the algorithm is correct and meets its purpose. 4) The evaluation ensures that the design of the programme is good and meets its purpose [40].			√	√	√	√
	11) Difficult to master the syntax of programming language	1) Students need to understand what they want to achieve and the process to achieve results. For example: a) The instructions for text display to specifically display the "Hello" text.			√		√	√

References	***Difficulties in programming***	***Justifications of mapping to computational thinking (CT) skills***	***Computational Thinking (CT) Skills***					
			Decomposition	***Abstraction***	***Pattern-Recognition***	***Algorithm***	***Logical Reasoning***	***Assessment or Evaluation***
		b) The instructions to enter inputs like scores. 2) Syntax is related to the programming language structure. Connecting with other concepts can assist students to learn syntax, such as a sentence must end with a period (.), while in programming, a statement ends with a semicolon (;). This is related to pattern-recognition skill. 3) Logical reasoning and evaluation skills help students to improve the ability to master syntax.						
	12) Difficult to understand the abstract concepts that involve the position of variables in computer memory	1) Analogies in daily life can help to understand the abstract concepts, such as an object with its content refers to the variable that holds its value. 2) The abstract concepts can also be conveyed using teaching aids, such as visualisation. For example, in computer memory, there is a variable that holds a value; the input entered would be held by the variable.			√		√	
Chan Mow [28]	13) Cognitive needs	1) CT skills refer to the cognitive process in problem solving [37].	√	√	√	√	√	√
Renumol, Jayaprakash, and Janakiram [41]	14) Cognitive difficulties	Same as the above problem (13)	√	√	√	√	√	√
Haberman and Muller [42]	15) Difficult to use the abstraction process	1) If problems are complex, decomposition skill is required to break down the problem into smaller parts; so, it would be easier to manage. Decomposition is a prerequisite for abstraction [14] when it comes to complex problems. The abstraction process to retrieve relevant information for each section is made after decomposition. 2) It is directly related to the need for abstraction skill. 3) Pattern-recognition skill can be helpful because the pattern-oriented instruction approach influences abstraction skill [27].	√	√	√			
Gomes & Mendes [26]	16) Cannot write a programme and develop an algorithm	1) Abstraction, decomposition, and pattern-recognition skills are best used as strategies to design solutions for the development of algorithms. It also requires logical thinking skill because an algorithm is a series of steps in the form that can be processed by a computer. 2) The study suggested emphasising the development of problem-solving skills among the novices. 3) Decomposition, abstraction, pattern-recognition, and algorithm skills are part of problem-solving skills. 4) Logical thinking is required to develop algorithms.	√	√	√	√	√	
Lahtinen, Ala-Mutka, and Jarvinen [25]	17) Difficult to understand the concepts	1) The study found the involvement of complex cognitive to understand without the phenomenon in daily life for			√	√	√	

References	***Difficulties in programming***	***Justifications of mapping to computational thinking (CT) skills***	***Computational Thinking (CT) Skills***					
			Decomposition	***Abstraction***	***Pattern-Recognition***	***Algorithm***	***Logical Reasoning***	***Assessment or Evaluation***
	of programming related to recursion, instruction, and abstract data	comparison. The study proposed appropriate design of teaching and learning materials to help students to master the concepts of programming through the comparison of concepts in life. Pattern recognition skill can help the students to master the concepts of programming that include: a) Recursive functions can be demonstrated through the concept of reuse, such as factorial or other situations in life. b) Demonstrates the use of pointers through the phenomena in daily life that can represent the concepts of programming c) Use other representations in daily life to illustrate abstract data 2) The concepts of programming are related to algorithm skill. 3) Logical thinking is necessary to master the concepts of programming.						
	18) Difficult to develop programmes	1) Materials such as the instructions to convert the algorithms to programme codes are used as references for students. Students can refer to the examples to write the programme. It is related to pattern-recognition skill. 2) The development of programmes involves syntax and semantics. Students must understand the flow of algorithms and use logical thinking skill to develop and evaluate programmes.			√	√	√	
Robins, Rountree, and Rountree [4]	19) Lack of strategy to plan the solutions and design algorithms	1) Decomposition, abstraction, pattern-recognition, and algorithm skills are important problem-solving skills for programming. Students need to think logically to design the algorithms.	√	√	√	√	√	
	20) Implement algorithms and write programmes	1) A guide to translate algorithms into suitable programme codes as a reference for students to implement algorithms. Approaches such as reusing, modifying, or integrating the existing programmes are strategies to develop new programmes. It is related to pattern-recognition skill. 2) Students must understand the flow of algorithms before writing the programme to ensure that the solution design is logical and valid. 3) The development and evaluation of programmes require logical thinking and assessment skills.			√	√	√	√
	21) Evaluate programmes and tracking and fixing errors	1) Require logical reasoning to evaluate and debug the programme					√	√
Winslow [24]	22) Problem to combine syntax and semantics to	1) Syntax is related to the structure of programming languages whereas semantics relates to the logic or concepts of statements, expressions, or			√	√	√	√

References	***Difficulties in programming***	***Justifications of mapping to computational thinking (CT) skills***	***Computational Thinking (CT) Skills***					
			Decomposition	***Abstraction***	***Pattern-Recognition***	***Algorithm***	***Logical Reasoning***	***Assessment or Evaluation***
	produce a valid programme	programmes. 2) Students need to understand the flow of algorithms before writing the program because an algorithm is a symbolic language that describes the solution in the form of programme code. It deals with algorithm and logical reasoning skills. 3) Syntax relates to the closeness of mapping, which is a relationship of programming languages with the students' existing knowledge of the concepts used. It is related to pattern recognition. 4) The semantics relates to the logic or concept for statement, expression, or programme. Students need to evaluate whether the algorithm or programme is logic and meets its purpose. Evaluation skill is used to ensure that the programme performs well and is able to achieve its goal [40]. Relates to logical reasoning and evaluation skills.						
Du Boulay [3]	23) Cognitive needs of programming	1) CT skills refer to cognitive processes in problem-solving [37]. Six CT skills are deemed very relevant to support problem solving in programming.	√	√	√	√	√	√
	24) Syntax and semantics	1) Syntax refers to the programming language structure. It is related to the closeness of mapping, which is the relationship of programming language with the students' existing knowledge of the problem or the concept that students want to learn. This is related to pattern-recognition skill. 2) The semantics relates to the logic or concept for statement, expression, or program are logic and valid. This stage uses logical reasoning and evaluation skills to evaluate an algorithm or programme.			√	√	√	√
	25) Lack of support skills (pragmatic)	1) Pragmatics refers to the practical aspects of how language features can be used to achieve multiple objectives. This is related to the logical reasoning skills. 2) Evaluation skill is also required to determine whether the programme is written in line with its objectives. The study then concluded that students need to learn the skills of how to determine, develop, test, and debug using the available tools.					√	√
	26) Orientation	1) The need to understand its uses, the types of problems that can be solved, and its advantages in programming 2) It is difficult for students to identify the terms of the programme, the actual process needed, and its usefulness. Students can be guided by programming-oriented contexts used in real life. It is directly related to pattern-recognition, logical reasoning and evaluation skills.			√		√	√

Overall, the difficulties in learning programming can be categorised as cognitive difficulties, difficulties in designing solutions, difficulties in developing algorithms, difficulties in writing and evaluating programmes, difficulties in combining syntax and semantics, difficulties related to the concepts of programming, and limited programming skills. The cognitive difficulties are related to difficulties in completing complex tasks, cognitive programming needs, lack of ability in mathematics, and less computational thinking. With limited cognitive abilities, it would also be difficult to design solutions. Difficulties in designing solutions refer to difficulties in using the abstraction process and designing solutions. As a result, it would be difficult to develop an algorithm as well.

Besides that, there are also the lack of strategies to design algorithms, difficult to understand the concepts of programming structure and programme design. Such scenario would inevitably lead to difficulties in writing and evaluating programmes. There are also other lack of strategies in implementing algorithms, difficulties in combining syntax and semantics (to produce a complete programme), and difficulties in evaluating programmes, debugging error, and tracking and correcting the errors. Moreover, learners who are not familiar with syntax would not be able to master semantics. Apart from the strategies to plan and develop algorithms, understanding the concepts of programming, which are typically related to recursive, pointer, abstract data, and mental models to implement programs, is also important. Last but not least, there are also the lack of support skills and orientation difficulties when it comes to the difficulties in learning programming.

B. *Mapping Validation by Experts*

These initial findings were then validated by the appointed experts. In this case, the mapping validation by experts referred to the expert evaluation on the CT skills in relation to the difficulties in learning programming. The mapping method in Fig. 1 was applied. The tabulated justifications of mapping to CT skills in Table III were reviewed by the experts. These experts were required to indicate their agreement to the established categories in the table.

Overall, the experts agreed on the mapped CT skills. There were additional expert reviews provided. For instance, students who are capable in mathematics may still encounter difficulty in mastering the logic of programming. The experts argued that the difficulties to come up with algorithms and programs are related to logical reasoning skill instead. In addition, certain students may not be able to come up with the solution because they evaluate the problem as a single, whole problem, rather than assessing the problem in different stages or processes. Such scenario demonstrates the students' incapability in problem-solving strategies, such as formulating problems to understand and design solutions and relate them to the existing knowledge and experience.

C. *Conducting a Survey*

Following that, an instrument was developed. In general, the instrument included a list of statements on the difficulties in learning programming. Considering the purpose of this study, the survey items aimed to measure all six identified CT skills, which were abstraction, decomposition, pattern-recognition, algorithm, logical reasoning, and evaluation skills. The details of the items for each construct are presented in Table II.

A total of 17 items were developed for each construct. A five-point Likert scale was employed. Likewise, these items were verified by experts before the actual data collection. Table III presents the level of measurement scale to determine the mean score of each construct.

A survey that involved instructors was then conducted to gather empirical data on the students' difficulties in learning programming. This study focused on programming lecturers from the Matriculation Division, Ministry of Education Malaysia. A total of 32 respondents participated in the survey. The survey aimed to identify the difficulties in learning programming based on the perspectives of these instructors. Through this survey, the need for CT skills to overcome the difficulties in learning programming can be identified to serve as a guide for the instructors in the teaching and learning process in programming.

TABLE II. DETAILS OF ITEMS FOR EACH CONSTRUCT

Construct	***Description of items***
Abstraction	Items were intended to identify the students' difficulties in understanding and formulating problems as well as identifying relevant information.
Decomposition	Items were intended to identify students' difficulties to decompose a problem that involves several processes in parts, so that it can be solved by section, as part of the problem-solving strategies.
Pattern-recognition	Items were intended to identify students' difficulties in integrating the existing knowledge and experience as problem-solving strategies.
Algorithm	Items were intended to identify students' difficulties in developing algorithms as well as their consequences if fail to create the algorithm.
Logical reasoning	Items were intended to identify students' difficulties in thinking logically by identifying and explaining the reasons behind the solution.
Evaluation	Items were intended to identify students' difficulties in evaluating the solution, whether the solution is suitable and meets its purpose.

TABLE III. LEVEL OF MEASUREMENT SCALE BASED ON MEAN SCORE

Mean score	***Level***
1.00 – 2.33	Low
2.34 – 3.67	Average
3.68 – 5.00	High

(Source: Landell, 1977) [43]

IV. ANALYSIS AND FINDINGS

There are two analysis and findings in this study. First, mapping the difficulties in learning programming with computational thinking skills. Second, survey among instructors on the students' difficulties in learning programming as well as the needs of CT skills. For the first analysis, the results of mapping help to identify the CT skills which is relevant to the difficulties in learning programming.

Some of the comments from experts described the problems encountered in programming learning. These results are useful for instructors to understand students' need in learning programming, helps to plan teaching and learning approaches or strategies and also in planning teaching materials and exercises.

Survey among instructors aim to identify students' difficulties in learning programming that in tandem with the needs of CT skills. Due to this purpose, data were analysed descriptively to determine the mean score of each item and construct, specifically to identify the level of need for each skill to overcome the difficulties in learning programming. As shown in Table IV, all items and constructs recorded high mean scores. The obtained results demonstrated that the difficulties in learning programming that were identified from literature are issues for learners. In addition, the results indirectly demonstrated the need for CT in teaching and learning programming. Algorithm skill recorded the highest mean score. This skill is useful for designing solutions by creating algorithms before writing the programs. Failure to develop algorithms may cause difficulty while coding. In addition, decomposition skill also showed high mean scores. This construct refers to the problems in managing large and complex problems. Due to this situation, it may cause problem to develop the algorithms. Based on these results, students need more problem solving practices to design the solutions. Hence, instructor must expose students with vary type of problems that require several processes to raise the strategies in problem solving. In conclusion, both of analysis and findings are useful to plan the strategy and approach in teaching and learning as well as to overcome the difficulties that commonly faced by students in programming.

TABLE IV. MEAN SCORE OF NEED FOR COMPUTATIONAL THINKING SKILLS IN TEACHING AND LEARNING PROGRAMMING

Construct	***Item***	***Mean score of items***	***Mean score of constructs***
Pattern-recognition	1	4.00	4.03
	2	4.06	
	3	4.03	
Decomposition	4	4.03	4.30
	5	4.34	
	6	4.53	
Abstraction	7	3.90	4.17
	8	4.22	
	9	4.38	
Algorithm	10	4.31	4.38
	11	4.41	
	12	4.41	
Logical reasoning	13	3.84	3.87
	14	3.78	
	15	4.03	
	16	3.75	
	17	3.94	
Evaluation	This item is contained indirectly in other constructs.		

V. CONCLUSION

Programming requires cognitive ability and involves strategies in planning and solving problems. Focusing on that, this study aimed to examine the difficulties in learning programming among students and determine the need for CT skills among instructors. This study first reviewed 11 empirical papers and three review papers on the difficulties in learning programming. Most of the identified difficulties in the past studies were related to cognitive needs, ability to plan solutions, difficulties in developing algorithms, and difficulties in writing and evaluating program. These identified difficulties were then mapped to the appropriate CT skills, which were validated by the experts. Following that, the items for each construct were developed for the survey. The survey specifically involved 32 instructors to gather empirical data on the difficulties in learning programming among students at the pre-university level. Based on the survey results, the identified difficulties in learning programming are clear among students today. Additionally, this directly demonstrated the need for CT skills in the teaching and learning process of programming. CT skills with appropriate approach or activities should be applied to guide students through real problems. The outcomes of mapping and survey were expected to contribute to the design of the problem-solving model and strategies in programming using CT skills, which can serve as a guide for instructors.

ACKNOWLEDGMENT

This study was funded by the UKM Research Grant (GUP-2018-155). We would also like to acknowledge the Matriculation Division, Ministry of Education Malaysia for their assistance and contribution for this study.

REFERENCES

[1] Vassilev, T. I. 2015. An Approach to Teaching Introductory Programming for IT Professionals Using Games. International Journal of Human Capital and Information Technology Professionals 6(1): 26–38.

[2] Stuikys, V. and Burbaite, R., 2018. Smart STEM-Driven Computer Science Education: Theory, Methodology and Robot-based Practices. Springer.

[3] Du Boulay, B. 1986. Some Difficulties of Learning to Program. Journal of Educational Computing Research 2(1): 57–73.

[4] Robins, A., Rountree, J. & Rountree, N. 2003. Learning and Teaching Programming: A Review and Discussion. Computer Science Education 13(2): 137–172.

[5] Yassine, A., Chenouni, D., Berrada, M. & Tahiri, A. 2017. International journal of emerging technologies in learning. International Journal of Emerging Technologies in Learning (iJET) 12(03): 110–127. Retrieved from http://online-journals.org/index.php/i-jet/article/view/6476.

[6] Watson, C. and Li, F.W., 2014, June. Failure rates in introductory programming revisited. In Proceedings of the 2014 conference on Innovation & technology in computer science education (pp. 39-44).

[7] Ibrahim, R., Rahim, N.Z.A., Ten, D.W.H., Yusoff, R., Maarop, N. and Yaacob, S., 2018. Student‟ s Opinions on Online Educational Games for Learning Programming Introductory. International Journal of Advanced Computer Science and Applications, 9(6), pp.332-340.

[8] Vieira, C., Yan, J. and Magana, A.J., 2015. Exploring design characteristics of worked examples to support programming and algorithm design. Journal of Computational Science Education, 6(1), pp.2-15.

[9] Margulieux, L.E. and Catrambone, R., 2016. Improving problem solving with subgoal labels in expository text and worked examples. Learning and Instruction, 42, pp.58-71.

[10] Jalani, N. H. & Sern, L. C. 2015. The Example-Problem-Based Learning Model: Applying Cognitive Load Theory. Procedia - Social and Behavioral Sciences 195: 872–880.

[11] Papert, S. 1980. Mindstorms; Children, Computers and Powerful Ideas. New York: Basic Book.

[12] Papert, S. and Harel, I., 1991. Situating constructionism. Constructionism, 36(2), pp.1-11.

[13] Wing, J. M. 2006. Computational thinking. Communications of the ACM 49(3): 33.

[14] Selby, C. 2015. Relationships: Computational Thinking, Pedagogy of Programming, and Bloom's Taxonomy. Proceedings of the Workshop in Primary and Secondary Computing Education 80–87.

[15] Estapa, A., Hutchison, A. & Nadolny, L. 2018. Recommendations to support computational thinking in the elementary classroom. International Technology and Engineering Educators Association. Retrieved from https://www.iteea.org/File.aspx?id=123563&v=25610bf

[16] Yadav, A., Gretter, S., Good, J. & Mclean, T. 2017. Computational Thinking in Teacher Education (November).

[17] Czerkawski, B. C. & Lyman, E. W. 2015. Exploring Issues About Computational Thinking in Higher Education. TechTrends 59(2): 57–65.

[18] Wing, J. M. 2010. Computational Thinking: What and Why? (November): 1–6.

[19] Shute, V. J., Sun, C. & Asbell-Clarke, J. 2017. Demystifying computational thinking. Educational Research Review 22: 142–158.

[20] Yadav, A., Hong, H. and Stephenson, C., 2016. Computational thinking for all: pedagogical approaches to embedding 21st century problem solving in K-12 classrooms. TechTrends, 60(6), pp.565-568.

[21] Bocconi, S., Chioccariello, A., Dettori, G., Ferrari, A., Engelhardt, K., Kampylis, P. & Punie, Y. 2016. Developing Computational Thinking: Approaches and Orientations in K-12 Education. Proceedings EdMedia 2016.

[22] Gretter, S. & Yadav, A. 2016. Computational Thinking and Media & Information Literacy: An Integrated Approach to Teaching Twenty-First Century Skills. TechTrends 60(5): 510–516.

[23] Kong, S. 2016. A framework of curriculum design for computational thinking development in K-12 education. Journal of Computers in Education 3(4): 377–394.

[24] Winslow, L. E. 1996. Programming Pedagogy --A Psychological Overview. ACM SIGCSE Bulletin 28(3): 17–22.

[25] Lahtinen, E., Ala-Mutka, K. & Jarvinen, H.-M. 2005. A study of the difficulties of novice programmers. ACM SIGCSE Bulletin 37(3): 14.

[26] Gomes, A. & Mendes, A. J. N. 2007. Learning to program-difficulties and solutions. International Conference on Engineering Education 1–5. Retrieved from http://ineer.org/Events/ICEE2007/papers/411.pdf

[27] Muller, O. & Haberman, B. 2008. Supporting abstraction processes in problem solving through pattern-oriented instruction. Computer Science Education 18(788840272): 187–212.

[28] Chan Mow, I. T. 2008. Issues and difficulties in teaching novice computer programming. Innovative Techniques in Instruction Technology, E-Learning, E-Assessment, and Education 199–204.

[29] Papadopoulos, Y. & Tegos, S. 2012. Using microworlds to introduce programming to novices. Proceedings of the 2012 16th Panhellenic Conference on Informatics, PCI 2012 180–185.

[30] Siti Rosminah, M. D. & Ahmad Zamzuri, M. A. 2012. Difficulties in learning programming: Views of students. 1st International Conference on Current Issues in Education (ICCIE 2012) (SEPTEMBER 2012): 74–79.

[31] Kwon, K. 2017. Student's misconception of programming reflected on problem-solving plans. International Journal of Computer Science Education in Schools 1(4): 14.

[32] Qian, Y. & Lehman, J. 2017. Students' Misconceptions and Other Difficulties in Introductory Programming. ACM Transactions on Computing Education 18(1): 1–24.

[33] Bundy, A. 2007. Edinburgh Research Explorer Computational Thinking is Pervasive Computational Thinking is Pervasive 1(2): 2–5.

[34] Witherspoon, E.B., Higashi, R.M., Schunn, C.D., Baehr, E.C. and Shoop, R., 2017. Developing computational thinking through a virtual robotics programming curriculum. ACM Transactions on Computing Education (TOCE), 18(1), pp.1-20.

[35] Lye, S. Y. & Koh, J. H. L. 2014. Review on teaching and learning of computational thinking through programming: What is next for K-12? Computers in Human Behavior 41: 51–61.

[36] Lopez, A. R. & Garcia-Penalvo, F. J. 2016. Relationship of knowledge to learn in programming methodology and evaluation of computational thinking. Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality - TEEM '16 73–77.

[37] Roman-Gonzalez, M., Perez-Gonzalez, J. C. & Jimenez-Fernandez, C. 2017. Which cognitive abilities underlie computational thinking? Criterion validity of the Computational Thinking Test. Computers in Human Behavior 72: 678–691.

[38] Barefoot Computing. t.th. Computational thinking. http://barefootcas.org.uk/barefoot-primary-computing-resources/concepts/computational-thinking/.

[39] de Araujo, A. L. S. O., Andrade, W. L. & Guerrero, D. D. S. 2016. A Systematic Mapping Study on Assessing Computational Thinking Abilities. 2016 Ieee Frontiers in Education Conference (Fie) 1–9.

[40] Csizmadia, A., Curzon, P., Dorling, M., Humphreys, S., Ng, T., Selby, C. and Woollard, J., 2015. Computational thinking-A guide for teachers.

[41] Renumol, V., Jayaprakash, S. and Janakiram, D., 2009. Classification of cognitive difficulties of students to learn computer programming. Indian Institute of Technology, India, 12.

[42] Haberman, B. & Muller, O. 2008. Teaching abstraction to novices: Pattern-based and ADT-based problem-solving processes. Proceedings - Frontiers in Education Conference, FIE 7–12.

[43] Landell, K., 1997. Management by menu. London: Wilay and Sms Inc.

Adapted Lesk Algorithm based Word Sense Disambiguation using the Context Information

Manish Kumar[1]
Department of Information Technology, SBPDCL
Bihar, India

Prasenjit Mukherjee[2]
Manik Hendre[3], Manish Godse[4]
Dept. of Analytics and IT
Pune Institute of Business management, Pune, India

Baisakhi Chakraborty[5]
Dept. of Computer Science and Engineering, NIT
Durgapur, India

***Abstract*—The process of identifying the meaning of a polysemous word correctly from a given context is known as the Word Sense Disambiguation (WSD) in natural language processing (NLP). Adapted Lesk algorithm based system is proposed which makes use of knowledge based approach. This work utilizes WordNet as the knowledge source (lexical database). The proposed system has three units – Input query, Pre-Processing and WSD classifier. Task of input query is to take the inputs sentence (which is an unstructured query) from the user and render it to the pre-processing unit. Pre-processing unit will convert the received unstructured query into a structured query by adding some features such as Part of Speech (POS) tagging, grammatical identification (Subject, Verb, and Object) and this structured query is transferred to the WSD classifier. WSD classifier uniquely identifies the sense of the polysemous word using the context information of the query and the lexical database.**

Keywords—Word Sense Disambiguation; natural language processing; WordNet; context; machine translation

I. Introduction

Natural language processing (NLP) is the field of computational linguistics or artificial intelligence that is concerned with the interaction between computers and human (natural) languages [1]. Natural language processing plays the important role in communication between human and machine. Word sense disambiguation is an important area of natural language processing and its application is related to find out the correct sense of an ambiguous word that is being used in a sentence. Many supervised and unsupervised algorithms have been developed on word sense detection as in [2]. In this field, a computer system is programmed in such a way that it is able to process a query provided in natural language and determine its correct semantics (meaning). Query is provided in the form of a sentence or a paragraph or text document. The semantics of a sentence depends on the semantics of its constituent words which are the smallest units of a sentence. Most of the words used in natural languages are associated with multiple meanings and these meanings vary frequently with the change in the contexts. Word with more than one meanings or senses are called polysemous words in the field of natural language processing and creates the problem of sense ambiguity. This work proposes a system such that it can correctly identify the meaning of the word (s) for the given context (sentence). Often a polysemous word has different meanings in different contexts. For example, the English word “bank” is associated with multiple meanings: “A financial institution”, “slopping land besides the water body”, or “have faith or confidence in” and many more as referred in the Princeton WordNet 3.0 [3]. The process of identifying the meaning of a polysemous word correctly in a given context is known as Word Sense Disambiguation (WSD).

Example 1

C1: I went to the **bank** to withdraw some money.

C2: Kolkata is situated at the **bank** of river Hugli.

Example 2

C3: **Cricket** is type of game.

C4: **Cricket** is a type of Insect.

Clearly, the word “bank” has two different meanings in the contexts C1 and C2 which are “a financial bank” and “land besides the river or sea” respectively. Similarly the word Cricket has also two different meanings in the contexts C3 and C4 respectively.

In the work proposed, knowledge based approach has been used for sense disambiguation. Natural language applications are using word sense disambiguation that is essential part in semantic analysis. Many models have been developed in word sense disambiguation where word space model is an effective model. This model represent the context vectors and sense vector in word vector space. Vector space is an important component to sense of a word as in [4]. Disambiguation of word sense using knowledge based approach can be done in two ways: Overlap method and Graph method. One of the pioneer works in overlap method is Lesk algorithm [5] that counts common words between two glosses (word definitions) to identify correct sense in the context. Gloss plays the vital role in the Lesk algorithm, which expresses two types of information: information about set of entries of all possible meanings and contextual information of target word. For the given pair of words, Lesk algorithm extracts meanings from lexical database and selects that sense as final sense which has the maximum overlap/common words/ co-occurred words. Lesk algorithm adapted Oxford Advanced Learner’s Dictionary as lexical database (also called as sense inventory). After some years later, two variants of Lesk [6] [7] have been proposed – “*Simplified*” version of Lesk algorithm [8] and Adapted Lesk algorithm [6]. The adapted Lesk algorithm adapted WordNet as lexical database and used the semantic relationships defined in the WordNet such as Hypernym, Hyponyms, Troponym, Meronyms, etc. Both Algorithms outperform the Lesk one as proved by Vosilescue et al. [9].

The specific meaning determination of an ambiguous word according to the context is a main task of word sense disambiguation. Mohammad Shibli Kaysar et al. [10] have introduced a system that is based on Bengali word sense disambiguation. A FP-Growth algorithm and Apriori algorithm have been proposed by the Authors on Bengali word sense disambiguation. The proposed system has been tested and 80% good result has been generated from ambiguous words as in [10]. Word sense disambiguation in Hindi language is limited. Anidhya Athaiya et al. [11] have approached Hindi language based word sense disambiguation that is genetic algorithm based. The proposed window is dynamic and feature of this window is containing vague word with left and right expression. The possible senses of an ambiguous Hindi word can be extracted from Hindi WordNet that is created by the IIT, Bombay as in [11].

The proposed work is an extension of adapted Lesk algorithm. Glosses provide the key information since they express the meaning of the words. There shall be two glosses, one corresponding to the target word, other corresponding to context word. In Lesk algorithm [5] and its variants [6] [7], there has been comparison between different senses of target word and context word in word pairs. In proposed work, the main focus is to decrease the number of comparisons between the word pairs. This would result in performance efficiency in terms of reduced time complexity. A significant improvement has been proved in time efficiency.

II. Related Works

In computational linguistic, Word sense disambiguation (WSD) is the ability to identify the correct meaning of words in context [12]. Contents on Internet are growing rapidly where existing sentences are containing ambiguous words. Removal of ambiguity from sentences that are containing ambiguous words is called word sense disambiguation as in [13]. In global word sense disambiguation, the shotgunWSD is a one of the best algorithm that is unsupervised and knowledge-based. ShotgunWSD has been developed from shotgun sequencing technique that is broadly applied in genome sequencing approach. The ShotgunWSD algorithm applies for word sense disambiguation at document level where it has three phases. The brute force algorithm is applied on short context window in first phase. In second phase, the local sense configurations are assembled by the prefix and suffix matching into the composite configurations where resulting configurations are ranked and sense of each word is detected based on majority voting as on [14]. WSD is a very common problem in the field of natural language processing (NLP). The WSD approaches used till date lies in the following two categories: Knowledge-based approach and corpus-based approach. Knowledge based approaches depends on the availability of knowledge sources such as thesaurus or dictionary or lexical databases (wordnet, BabelNet) to perform disambiguation. Knowledge-based approach uses two types of methods: Overlap method and Graph method. Corpus-based approach uses sense tagged corpus (supervised approach) and sense untagged corpus (unsupervised approach). The first noticeable work of knowledge based approach is by the Lesk [5]. The basic idea used in this algorithm is the sense definition or gloss of the word. In this algorithm, the gloss of the target word is compared with the gloss of all other context words and a score is calculated. Score is defined as the number of common words between the gloss of the target word and the gloss of the context words. Sense with the maximum score is the winner and is assigned as the final sense for the target word in the given context. There are several variants of Lesk algorithm have been proposed [6] [7] [8] [9] [15]. Kilgarriff and Rosenzweig [7] proposed a *Simplified version of Lesk algorithm*. They disambiguated each word individually and it results in the decrease in number of comparison of word pairs. Banerjee and Pederson [6] proposed Adapted Lesk Algorithm for WSD using WordNet. In this algorithm, authors have used the WordNet as lexical resources and they explored the concept of semantic relations defined in the WordNet such as Hypernym, hyponym, meronym, Troponym, Holonym and attributes of each word glosses. In the next work, Banerjee and Pederson [6] presented a new algorithm to measure the semantic relatedness between concepts: *"Extended Gloss Overlaps as a Measure of Semantic Relatedness"*. The measure was number of words matches between the definitions of senses (glosses). They extended the gloss of the concept by incorporating the gloss of other related concepts as defined in the WordNet concept network. A relative evaluation was performed on the variants of Lesk's algorithm by vasilescu et al. [9]. they found that the variants of the Lesk algorithm outperformed the original Lesk algorithm. Baldwin et al. [15] suggested a new algorithm of Machine Readable Dictionary (MRD) based WSD using definition extension and ontology induction. They have used the basic idea of original Lesk [5] and Adapted Lesk algorithm [6]. They experimented over the Hinoki Sense bank and the Japanese Senseval-2 datasets and they found that sense-sensitive definition extension over semantic relations defined in WordNet, integrated with definition extension and word tokenization leads to WSD accuracy above both unsupervised and supervised baselines. Wang and Hirst [16] proposed a method to measure WSD using Naive Bayes similarity. In this method, they replace the overlap mechanism of the Lesk [5] with a general-purpose Naive Bayes model applying the maximum likelihood probability approximate. Brody and Lapata [17] presented an unsupervised approach to determine WSD: Good Neighbours Make Good Senses using distributional similarity. They applied distributional similarity to identify similar words and prepare a sense tagged training dataset without human efforts which is further used to train a standard supervised classifier for doing sense disambiguation. They have adapted Senseval-2 and Senseval-3 dataset for the experiment and got remarkable improvements over state-of-the-art unsupervised methods of WSD. Khapra et al. [18] proposed Bilingual Bootstrapping method for WSD. Considered the bilingual language setup, where the languages under consideration are having fewer amounts of seed data but have the sufficient amount of untagged data. Their idea of tagging the untagged data of one language using the seed data of other language and vice-versa is solely based on bootstrapping method using parameter projection. They use Hindi and Marathi as language pair for their experiment. Khapra et al. [19] proposed a domain specific iterative WSD method for multilingual setting. They considered Hindi, English and Marathi languages for lingual setting. This method is completely dependent on the dominant senses of words (can be nouns, adjectives and adverbs) in the specific domain to accomplish disambiguation. An overall accuracy of 65% on F1-score was reported for all the three languages. Zhong et al. [20] introduced a new WSD method for free text. They utilized the idea of linear support vector machine (SVM) as classifier with some knowledge based features. Singh and Siddiqui [21] proposed an overlapping based WSD method for Hindi. They examined the effect of the removal of stop word, stemming and context window of different sizes and they noticed an improvement of 9.24% and 12.68% in precision and recall respectively. Heyan et al. [1] suggested a new method of unsupervised WSD using collaborative technique. In this work, they utilized the within-sentence relationship (ambiguous sentence) as well as cross sentence relationship (neighbour sentence). The graph-based ranking algorithm is used to perform the disambiguation task. Navigli & Lapata [3] proposed a graph based method for unsupervised WSD. They utilized measures of graph connectivity to find out the most important node (sense) in the graph. They also evaluated the role of lexicon selection and sense inventory as it helps in determining the structure of sub-graph of graph. They used the SemCor dataset for the experiment and show that the degree centrality provides best results compared to the other well known WSD technique such as PageRank, Betweenness Centrality, HITS and Key Player Problem. Basile et al. [22] proposed a WSD algorithm using distributional semantic model. This work relies on the variants [6] [8] of Lesk algorithm. Their approach solely depends on the word similarity function defined over semantic space i.e. they did not use direct matching of words but they used cosine similarity function to get score of overlap. They performed the experiment over SemEval-2013 dataset and adapted BabelNet as lexical database. The Semantic analysis is a crucial part of NLP systems such as information retrieval, data mining, and machine translation. In [23], Authors have discussed about the word vector space model that has been extended to reflect more accurate meaning in context vectors. In [24], Authors have elaborated about the Word Sense Disambiguation in Bengali Language. The Induction technique has been used in first phase of this system where second phase is Word Sense Disambiguation which is developed by the use of Semantic Similarity Measure. ShotgunWSD [25] is a recent algorithm of The Global word sense disambiguation (WSD) which is unsupervised and knowledge-based algorithm. The algorithm has been developed from the Shotgun sequencing technique. The ShotgunWSD contains three phases. The first phase is a brute-force algorithm, the second phase assembles local sense configurations to longer composite configurations and third phase is related to chosen of the sense of each word which is based on a majority voting scheme.

III. Architecture

Initially, the query (in English) is provided by the user in the form of a sentence or a paragraph or a text document. User provided query is an unstructured text (not having any features such as part-of speech, stemmed form of words, etc. attached with it). Each query contains only a single target word (a polysemous word). Target word can be of any type of WordNet word (Noun, Verb, Adjective, and Adverb). A query is provided by the user in English language, must follow the rule of English grammar that is (S + V + O) and must follow the following production rule:

$S \rightarrow NP + VP$

$NP \rightarrow NN\ /\ PRN\ /\ (DET + NN)$

$VP \rightarrow (VB + NP)\ /\ (VB + PP)$

$PP \rightarrow (TO + NP)\ /\ (TO + VP)$

The Subject (S), Verb (V) and Object (O) of any query can be represented as follows,

$S \in NP,\ V \in VB$

$O \in \{VP - \{VB\}\} \Rightarrow \{O \in NP\ /\ O \in PP\}$

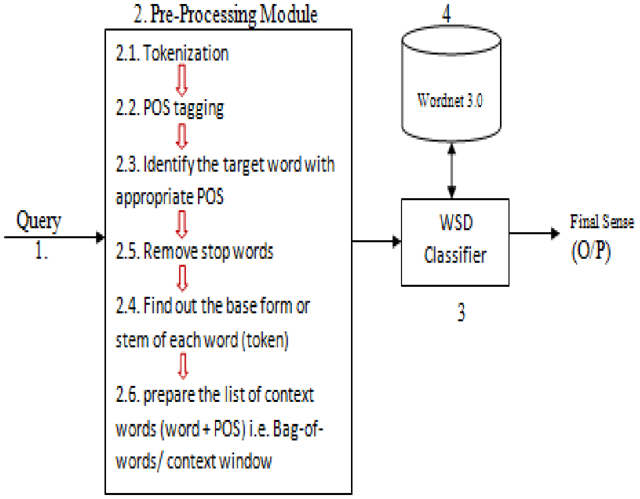

2. Pre-Processing Module
2.1. Tokenization
2.2. POS tagging
2.3. Identify the target word with appropriate POS
2.5. Remove stop words
2.4. Find out the base form or stem of each word (token)
2.6. prepare the list of context words (word + POS) i.e. Bag-of-words/ context window
4
Wordnet 3.0
Query
1.
WSD Classifier
3
Final Sense
(O/P)

Fig 1. Modular Architecture of WSD System.

The architecture of proposed work has been given in Fig. 1. Target word used in the query can be of any type of WordNet word and according to its type context window is prepared which is discussed in following cases:

1) Case 1: w_t can be of any type of WordNet word except verb

In this case, target word can be either noun, Adjective or Adverb. If there are n number of words appears to the left and to the right of the target word then context window is prepared of size (2*n + 1). In the proposed system, only those pairs of words are used that contain target word. So total numbers of words pairs possible are 2n.

2) Case2: when target word is a Verb (W_T = VB & VB ∈VP)

If the target word is verb and is part of VB in the given query then its dependency is more on the object of the query so left side words are ignored by the proposed system. Total numbers of word pairs formed are n as size of the context window is (n+1).

3) Case3: when target word is a Verb (W_T = VB & VB ∊PP)

If the target word is a verb and it is preceded by a preposition then it is the part of PP. In this case, subject and verb part of VP are ignored.

Identification of WordNet type of the target word is performed by POS tagger. The full discussion of POS tagger is explained in subsection2 of pre-processing module of the proposed system.

A. Pre-Processing Module

Pre-processing module is responsible for taking user query (instance/example containing target word) and converts this query to a structured text. To convert the input query to structured text following steps are performed by the pre-processing module:

1) Tokenization: It is the process of breaking the input query into individual words. Each word is known as token.

2) POS Tagging: This is a process to identify the correct part of speech for each word of the input query. The Adapted standard POS tagger is for annotating the input query. There are many abbreviations are defined to tag words, but some of them have been used that are NN-Noun, NNP-proper Noun, VB - verb, DT - determiners, TO – preposition, etc.

3) Target word is identified and is attached with its proper POS tag.

4) Stop words: the word which appears frequently in the context but the meaning of the context doesn’t depend on that word is considered as stop word. In this work, stop word includes all non-WordNet words and auxiliary verbs. If word is a stop word then it will be removed from the context.

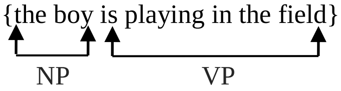
5) Stemming: this is a process of converting each word into its original (base) form.

6) Lastly context window or Bags- of- words is prepared. It contains all the context words including target words.

B. Example:

“The boy is playing in the field.”

Tokenization: {The, boy, is, playing, in, the, field}
POS Tagging: {The/DT, boy/NN, is/VBD, playing/VB, in/TO, the/DT, field/NN}
Chunking: {the boy is playing in the field}



Stemming: {The/DT, boy/NN, be/VBD, play/VB, in/TO, the/DT, field/NN}
Stop words: {The, is (be), in}
Context window: {boy/NN, play/VB, field/NN}
After the completion of pre-processing, the bag-of-words contains the word attached with their features like attachment of POS tag. These texts are called structured text that will be delivered to WSD classifier.

C. WSD Classifier

Word sense disambiguation classifier is the last and core module of the proposed system. This module is responsible for the following tasks:

1) Search WordNet: the first task of WSD classifier is to search for the senses of each word $\mathbf{w_i}$ of the context window and retrieve those senses. To retrieve senses, WSD classifier interacts with WordNet which is used as the lexical database. As the type of the word is given in the context window so only matching type of senses are retrieved. For an example, for the word ‘play/NN’ only noun type of senses is retrieved from the WordNet. After retrieving the senses for all the words of context window, the separate lists of senses have been prepared for the target word and other context words.

2) Score calculation: score is the number of words common between the two glosses. There shall be two glosses, one corresponding to the target word, other corresponding to context word. The methodology for score calculation in given in the algorithm 1.

D. Algorithm

```
Score_calculation ($S_t$ , $S_{cw}$ )
  Input: List of senses definitions (glosses) of Target word
and Lists of senses of context words
  Output: Any one sense of Target word
  SC ← Φ, $SC^C$ ← Φ, $SC^I$ ← Φ
    For (i = 1 to | $S_t$ |)
        For (c = 1 to | CW |)
            For ( j = 1 to $N_c^S$)
                SC ← SC U {overlap ($g_t^i$ , $g_c^j$ )}
            End for
            $SC^C$ ← $SC^C$ U {Maximum (SC)}
        End for
        $SC^I$ ← $SC^I$ U {maximum ($SC^C$)}
    End for
  f = argmax $SC^I$
Return f
```

Some notations are used in the algorithm1 which are explained below:
S_t: List of all sense definitions (glosses) of the target words
S_{cw}: Lists of sense definitions (glosses) of context words.
SC: set of scores calculate for all glosses of any one context word.
SC^C: Set of maximum scores obtained for each context words for ith sense of target word.
SC^I: Sets of maximum score obtained for each sense of target word from all the context words.
f: sense number of the target word which has maximum score and this sense is winner.
Score_calculation () method return the sense number of the target word which having maximum score for all the context words and all other senses of target word. This sense is the correct sense for that context.

IV. Advantages of Proposed System

1) Reduction in Number of Comparison

In the adapted Lesk algorithm, authors have used the concept of overlapping mechanism on the glosses of all the possible pairs of words of the context window where size of the window is equal to the sum of the left and right neighbouring words of the target word and the target word itself. Let 'n' is the number of words to the left and right of the target word then window size is equal to (2*n + 1). In the adapted Lesk algorithm, author have considered all the possible pairs of context words ($^{2n+1}C_2$), but in the proposed work considering only those pairs of words which are having target word. The proposed system is trying to find out the correct sense from the list of senses of the target word, only those pairs can provide the useful information which is containing target word. Total numbers of word pairs possible in the proposed work are 2n.

A/C to adapted Lesk Algorithm,
Total no. of pairs of context window = $^{2n+1}C_2 = (2n+1)*2n/2 = n*(2n+1) \Rightarrow O(n^2)$
But A/C to our approach,
Case1: Total no of pairs of context window = $1 * 2n = 2n \Rightarrow O(n)$
Case2 & case3: Total no. of pairs of context window = $1*n = n \Rightarrow O(n)$
Note: In cases 2 and case 3, the subject part and subject and verb part of the query have been ignored respectively. That means the words have been ignored to the left of the target word so that the size of the window gets decreased to n+1. Total numbers of pairs possible are n.
The proposed work is using less number of word pairs as compared to the adapted Lesk algorithm and so it takes less time to compare the glosses. Word pairs are in O (n) in the proposed work whereas O (n^2) in the adapted Lesk algorithm.

2) Use of POS tagger:

Let,

$S = \{s_1, s_2, s_3, ..., s_i, ..., s_N\}$; set of all senses of the target word '$\mathbf{w_t}$'

Total number of sense = $|S| = N$

$\mathbf{S_{pos} \subset S}$ where pos= {noun, verb}

$S_{noun} = \{s_1, s_2, s_{3,}....... s_{n1}\}$

$S_{verb} = \{s_1, s_2, s_3, ..., s_{n2}\}$

$|S_{noun}| = n1$, $|S_{verb}| = n2$

Therefore, $|S| = |S_{noun}| + |S_{verb}|$

⇨ $n1+n2 = N$

⇨ $n2=(N-n1)$

$S_i \in S$ is any ith sense of the target word w_t

a) Without using POS tagging:

Let us consider a target word '$\mathbf{w_t}$' in a given context C. Target word should be either Noun or Verb. If this word has total senses available in the WordNet is N. To determine the correct sense of the target word '$\mathbf{w_t}$', now, consider all N senses out of which only one will be the correct sense.

b) Using POS tagging:

To get the correct part-of-speech of the target word, The POS tagger has been used. The target word $\mathbf{w_t}$ must be either a noun or a verb.

Let,

Total sense of $\mathbf{w_t}$ as noun = n1

Total sense of $\mathbf{w_t}$ as verb = (N-n1)

Where n1<=N,

Note: Equality holds if the target word $\mathbf{w_t}$ has any single type of POS tag.

In this case, the type of POS of the target word $\mathbf{w_t}$. If $\mathbf{w_t}$ is a noun then consider only n1 senses otherwise (N-n1) senses to correctly identify the sense of the target word $\mathbf{w_t}$. The lesser number of senses (<N) have been used in both cases except for only one type of POS tag applied for the target word.

POS tagging is applied on all the words of the context window. So it saves a lot of time and space since less numbers of senses are used in the comparison. Pos tagging process can take some extra time but overall it performs better in respect of time. This is analyzed by us by doing several experiments.

Example: The boy is playing in the field.

In the above example, play (stemmed form of playing) is the target word as provided by the user. After POS tagging, the target word play is identified as a verb. So the probability of its meaning's dependency is more on the object of the query.

Query: Q = {*the boy is playing in the field*}
|NP||VP|

NP = {the boy}

VP = {is playing in the field} ⇒ VB = {is playing}; PP = {in the field}

S = the boy, V = is Playing and O = in the field

After stemming and removing stop words final schema of the query is

S = {boy}, V = {play}, O = {field}

case1: If $W_T \neq VB$:

Context window (WC) =

Boy, play, field
-1 0 +1

Pairs: {(play, field), (play, boy)}

Case2: If $W_T = VB$

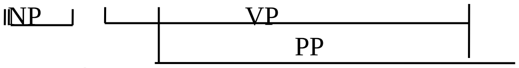
Context window (WC) =

Play, field
0 +1

Pairs: {(play, field)}

Case3: If $W_T = VB$ and $W_T \in PP$

Query: "*Shyam likes to play with emotions of people*."



NP: {Shyam/NN}

VP: {likes to play with emotions of people}

⇒VB: {likes} && PP: {to **play** with emotions of people}

After removing stop words and stemming,

S = {shyam}, V= {likes}, O = {play with emotions of people}

Context window (WC) =

Play,	emotion,	people
0	+1	+2

Pairs: {(play, emotion), (play, people)}

V. Results and Discussion

The proposed WSD system is implemented on JAVA platform utilizing WordNet3.0 API along with Stanford POS tagger. Lemmatizer has been used to extract the base form of the word. The proposed system has been tested on 50 highly polysemous English words. These 50 polysemous words are taken into three categories - Nouns (20), Verbs (20), and Adjectives (10) and 2000 example sentences are defined for these words. These examples are taken from Wiktionary[1] and other sources of internet[2]. The stop word removal and lemmatization have been implemented which increase the number of overlaps.

The Precision (P), recall (R) and attempt (A) have been measured for the proposed system. The system gives correct output for 1330sentences out of 2000 sentences. The attempt of the system is 100%, so P and R are equal. The P, R, A value are given in Table I. Proposed model classified the query in either correct output or incorrect output. Number of incorrect result is same for both P & R. so P & R are same.

TABLE I. Possible Word Pairs in WSD Approaches

	R	P	A
Lesk	0.23	0.23	100%
A-Lesk	0.47	0.47	100%
M-Lesk	0.665	0.665	100%

Overall performance of the proposed system is better than Adapted Lesk algorithm.

VI. Future Work

In future, the focus will be on the selection of context bag in such a manner which improves the accuracy of system. We can also improve upon the accuracy obtained in case of the single occurrence of target word by working on a standard dataset and thereby comparing the results with that of obtained from other knowledge based approaches. There is also some scope of improvement in WSD as we can use babelNet which is a multilingual WordNet.

VII. Conclusion

In this paper word-sense disambiguation (WSD) problem in natural language processing is studied. WSD governs the process of identifying sense of a word or meaning which is used in a sentence, when the word has multiple meanings (polysemy). The analysis is done and results are compared for both single occurrence of target word as well as multiple (two) occurrences of target words. Experimental results shows that proposed modified lesk method gives 0.665 precision and recall which is higher than Lesk and Adaptive Lesk methods. Due to the improper selection of context bag, lesk and adpative lesk algorithms gives poor results. The proposed method reduces the number of word pair comparisons as compared with the adaptive Lesk Algorithm. Proposed method needs O (n) word pairs as compared with O (n^2) required by the adaptive Lesk algorithm.

References

[1] H. Heyan, Y. Zhizhuo, and J. Ping, "Unsupervised word sense disambiguation using neighborhood knowledge", In 25th Pacific Asia Conference on Language, Information and Computation, pp. 333-342, 2011.

[2] H. Walia, A. Rana and V. Kansal, "A Supervised Approach on Gurmukhi Word Sense Disambiguation Using K-NN Method," 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, pp. 743-746, 2018.

[3] R. Navigli, and M. Lapata, "An experimental study of graph connectivity for unsupervised word sense disambiguation", IEEE transactions on pattern analysis and machine intelligence 32, no. 4, pp. 678-692, 2010.

[4] M. Y. Kang, T. H. Min and J. S. Lee, "Sense Space for Word Sense Disambiguation," 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), Shanghai, pp. 669-672, 2018.

[5] M. Lesk, "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone", In Proceedings of the 5th Annual International Conference on Systems Documentation, SIGDOC '86, pp. 24-26, New York, NY, USA. ACM, 1986.

[6] S. Banerjee and T. Pedersen. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In lexander Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, volume 2276 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, pp. 136–145, 2002.

[7] A.Kilgarriff, and J. Rosenzweig, "Framework and results for English SENSEVAL", Computers and the Humanities, 34(1), pp.15-48, 2000.

[8] S. Banerjee and T. Pedersen, "Extended Gloss Overlaps as a Measure of Semantic Relatedness", Ijcai, vol. 3, pp. 805-810, 2003.

[9] F.Vasilescu, P. Langlais, and G. Lapalme. "Evaluating Variants of the Lesk Approach for Disambiguating Words". In Proceedings of the 4th Conference on Language Resources and Evaluation (LREC), pp. 633–636, 2004.

[10] M. S. Kaysar, M. A. B. Khaled, M. Hasan and M. I. Khan, "Word Sense Disambiguation of Bengali Words using FP-Growth Algorithm," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, pp. 1-5, 2019.

[11] A.Athaiya, D. Modi and G. Pareek, "A Genetic Algorithm Based Approach for Hindi Word Sense Disambiguation," 2018 3rd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, pp. 11-14, 2018.

[12] R. Navigli, "Word sense disambiguation: A survey", ACM Computing Surveys (CSUR) 41, no. 2, 2009.

[13] E.Faisal, F. Nurifan and R. Sarno, "Word Sense Disambiguation in Bahasa Indonesia Using SVM", 2018 International Seminar on Application for Technology of Information and Communication, Semarang, pp. 239-243, 2018.

[14] M. Butnaru and R. T. Ionescu, "ShotgunWSD 2.0: An Improved Algorithm for Global Word Sense Disambiguation," in IEEE Access, vol. 7, pp. 120961-120975, 2019.

[15] I.T. Baldwin, S. Kim, F. Bond, S. Fujita, and D. Martinez. "A Reexamination of MRD-Based Word Sense Disambiguation", In Journal of ACM Transactions on Asian Language Information Processing (TALIP) Volume 9 Issue 1, Article No. 4, March, 2010.

[16] T.Wang, and G. Hirst, "Applying a Naive Bayes Similarity Measure to Word Sense Disambiguation", In ACL (2), pp. 531-537, 2014.

[17] S.Brody and M. Lapata, "Good Neighbors Make Good Senses: Exploiting Distributional Similarity for Unsupervised WSD", In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, United Kingdom , Volume 1, pp. 65–72, 2008.

[18] M. Khapra, M. Mitesh, S. Joshi, A. Chatterjee, and P. Bhattacharyya. "Together we can: Bilingual bootstrapping for WSD", In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 561-569, 2011.

[19] M. Khapra , P. Bhattacharyya, S. Chauhan, S. Nair, A. Sharma,"Domain specific iterative word sense Disambiguation in a multilingual setting", In Proceedings of International Conference on NLP (ICON 2008), Pune, India, Dec. 2008.

[20] Z. Zhong and H. T. Ng., "It makes sense: A wide-coverage word sense disambiguation system for free text", In Proceedings of the ACL 2010 System Demonstrations, pp. 78-83, 2010.

[21] S. Singh, and T. J. Siddiqui, "Evaluating effect of context window size, stemming and stop word removal on Hindi word sense disambiguation", In Information Retrieval & Knowledge Management (CAMP), IEEE, pp. 1-5, 2012.

[22] P. Basile, A. Caputo, and G. Semeraro, "An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model", In COLING, pp. 1591-1600, 2014.

[23] M. Y. Kang,T. H. Min and J. S. Lee, "Sense Space for Word Sense Disambiguation," 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), Shanghai, pp. 669-672, 2018.

[24] A. Sau, T. A. Amin, N. Barman and A. R. Pal, "Word Sense Disambiguation in Bengali Using Sense Induction," 2019 International Conference on Applied Machine Learning (ICAML), Bhubaneswar, India, pp. 170-174, 2019.

[25] A. M. Butnaru and R. T. Ionescu, "ShotgunWSD 2.0: An Improved Algorithm for Global Word Sense Disambiguation," in IEEE Access, vol. 7, pp. 120961-120975, 2019.

AUTHORS' PROFILE

Manish Kumar received M. Tech in Information Technology from National Institute of Technology, Durgapur, India, in 2015. He is working as an Asst. IT Manager in South Bihar Power Distribution Company Ltd., Bihar, India. His research interest includes Natural Language Processing, Database Management System, Knowledge Management System and Mathematical Analysis.

Prasenjit Mukherjee has 12 years of experience in academics and industry. He was a fulltime Ph.D. Research Scholar in Computer Science and Engineering in the area of Natural Language Processing from National Institute of Technology (NIT), Durgapur, India under the Visvesvaraya PhD Scheme from 2015 to 2019. Presently, He is working as a Data Scientist under Analytics and IT Department, Pune Institute of Business Management, Pune, Maharashtra, India.

Manik Hendre has 5 years of experience in academics and industry. He is a Ph.D. research scholar at the University of Pune. He is currently working as Data Scientist in RamanByte Pune. His research areas include Biometrics Image Processing, Machine learning and Data Analytics.

Dr. Manish Godse has 25 years of experience in academics and industry. He holds Ph.D. from Indian Institute of Technology, Bombay (IITB). He is currently working as an Industry Professor and IT Director in the PIBM, Pune in the area of Artificial Intelligence and Analytics. His research areas of interest include automation, machine learning, natural language processing and business analytics. He has multiple research papers indexed at IEEE, ELSEVIER, etc.

Dr. Baisakhi Chakraborty received the PhD. degree in 2011 from National Institute of Technology, Durgapur, India in Computer Science and Engineering. Her research interest includes knowledge systems, knowledge engineering and management, database systems, data mining, natural language processing and software engineering. She has several research scholars under her guidance. She has more than 60 international publications. She has a decade of industrial and 14 years of academic experience.

An Application of Zipf’s Law for Prose and Verse Corpora Neutrality for Hindi and Marathi Languages

Prafulla B. Bafna[1], Jatinderkumar R. Saini[2]
Symbiosis Institute of Computer Studies and Research
Symbiosis International Deemed University, Pune, India

***Abstract*—Availability of the text in different languages has become possible, as almost all websites have offered multilingual option. Hindi is considered as official language in one of the states of India. Hindi text analysis is dominated by the corpus of stories and poems. Before performing any text analysis token extraction is an important step and supports many applications like text summarization, categorizing text and so on. Token extraction is a part of Natural language processing (NLP). NLP includes many steps such as preprocessing the corpus, lemmatization and so on. In this paper the tokens are extracted by two methods and on two corpora. BaSa, a context-based term extraction technique having different NLP activities, e.g. Term Frequency Inverse Document Frequency (TF-IDF) and Zipf’s law are used to count and compare extracted tokens. Further token comparison between both of the methods is achieved. The corpus contains proses and verses of Hindi as well as the Marathi language. Common tokens from corpora of verses and proses of Marathi as well as Hindi are identified to prove that both of them behave same as per as NLP activities are concerned. The betterment of BaSa over Zipf’s law is proved. Hindi Corpus includes 820 stories and 710 poems and Marathi corpus includes 610 stories and 505 poems.**

Keywords*—*Marathi; NLP; Synset; Zipf’s law

I. Introduction

Hindi and Marathi languages are not only popular in the world but also are used as an official language in North India and Maharashtra, respectively [1]. So, abundant Hindi and Marathi text get generated day by day. To process this data NLP techniques along with machine learning algorithms are available in the literature. Generally, to analyze the behavior of algorithms, the corpus of Hindi or Marathi poems and stories is being used. Poems and stories are part of the literature. Stories and poems act as a guide to children about their behavior and manners [2-3] and connect with elders to interconnect ideas and visualize life’s opportunities. The use of rhyme and meter gives musical sense to the poetry, which is termed as literary elements whereas stories include a set of incidents and characters. Nouns, adjectives, adverbs are prominently used to construct a story or a poem [4].

NLP processing on this corpus is carried out after the collection of data and the creation of a corpus. There are three steps implemented on a corpus which are tokenization, noise removal and normalization [5-8]. Separating the text strings into smaller units is known as tokenization. Paragraphs can be tokenized into sentences and sentences can be tokenized into words [9-11].

Removal of noise or stop word removal is carried out after tokenization. Stop words are those which need to be deleted from the corpus to remove noise. These are the words, which are not important and increases attributes. Eg. ‘में’ (mei) in Hindi and ‘या’ (ya) in Marathi, that is in, punctuations, numbers, etc. The next step is normalization, stemming and lemmatization are part of normalization. It reduces the word to its base form. Lemmatization [12-15] is said to be more accurate than stemming. It reduces word to a meaningful form. E.g. Lemma of studies is study and stem is studi. Lemma uses morphological analysis. Stem removes inflectional ending only.

After lemmatization, generally, term frequency-inverse document frequency is used. It is based on a total number of terms present in the corpus. The importance of the term increases as its count is increased but it is offset by inverse document frequency. Terms present in almost all of the documents are ignored. TF-IDF measure is assigned to the significant terms in the corpus.

Zipf’s law is another measure to decide the significance of terms [16]. When applied to the language it states that the top 20% of the most frequently used words in a corpus large enough will make up 80% of it.

To make it clearer, say a novel contains 5000 different words. According to the rule, 80% of the novel will be the most frequently used 1000 words. It allows us to extract all terms/ words and states that the rank of a word is inversely related to its frequency. Mathematically, terms having frequency 40% of maximum frequency are significant. For e.g., If the maximum frequency of the term in the corpus is 100 the terms having frequency >=40 are significant.

In this paper statistical analysis of the tokens present in both the corpus is depicted visually along with the common tokens at each stage. It will provide guidelines to researchers working in the metalinguistic domain.

Corpora containing more than 1500 documents, more than 3 Million terms words are processed. Statistical visual analysis of the terms is carried out using BaSa [1]. Zipf’s law is applied to the same corpus and common tokens are identified. Similarly, Marathi corpus is created to apply BaSa and common tokens are extracted for both of the corpora, also Zipf’s law is applied on Marathi corpus common tokens are identified for Marathi corpus too. Finally, Common tokens extracted by Zipf’s law and BaSa are compared for both of the corpora. This research is unique because:

1) More than 3 Million terms are processed

2) Context-based token comparison with lemmatization and its visualization is done the first time

3) 4 types of corpora with multilinguistic context-based approach are processed

Processing proses and verses are one and the same with respect to NLP activities.

II. Background

India is a diverse country having around 23 different official languages and this has opened a wide area for natural language processing researchers. Indian language domains have lots of data accumulated in recent years and thus provided opportunities to mine this data.

A model is proposed for carrying out a sentiment analysis on Hindi tweets. It also focuses on the challenges of sentiment mining for Hindi tweets. The accuracy of the model is calculated [17].

Sentiment analysis [18-20] for Indian languages has become significant due to data present in Indian languages has expanded online and offline. The growth of Indian languages over a period in the area of sentiment mining is stated along with the taxonomy of Indian languages.

It will provide sources of datasets with annotation for linguistic analysis and suggest the appropriate technique for sentiment analysis in a specific domain.

Different types of stemming techniques for Indian and Non-Indian languages are explained. The algorithm is proposed to retrieve the set of Marathi documents based on the users' requirements. The rule-based approach is followed by stemming techniques, which always performs better Brute force. Stemmers are build using NLP techniques along with Dictionary-based algorithms. The stemmers allow encoding different language-related rules. These stemmers are suitable for a specific language. A text summary of Marathi documents is performed by extracting tokens present in the data. It is done by abstracting documents and using morphological rules of language. It reduces the time and effort invested in reading the documents [2][21].

Due to large text available on different applications like travel aggregator, google assistant, the need for text summarization is evolved across the period. Summarization gives an abstract view of data in fewer words without changing its meaning. Different challenges of text mining are explored such as context based analysis and so on.

Different Indian languages are explored by different researchers and NLP elements explored for each language are stated. Poetry corpus creation along with preprocessing of the corpus is achieved by Punjabi corpus and classifiers are executed. Diacritic extraction methods are used for the Gujarati language along with information retrieval, stop word identification and classification and machine translation. List of stop word its analysis building dictionary, constituency mapping, development of lemmatizers and morphological analysis are developed in Sanskrit [22]. Metadata is generated related to poetry and Hindi text analysis was performed. Stemming is used to improve the performance of the algorithm and it is a preprocessing technique. It removes tagging of the word and reduces it and used in information retrieval.

The sensitivity performance of negative news articles is implemented. News articles are classified as positive, negative and neutral. The articles formed different domains that are sports, politics and so on. Local administration cannot take action against such news. Some news may be urgent to treat can be focused by a proposed approach. TF-IDF is used on unigrams and bigrams of 1000 news collected from websites and performance of the classifier were evaluated [3][21].

To predict about the occurrences of the terms, Zipf's Law is used as base-line rule. Word's frequency decides its role in the entire corpus. The semantic influence of a word and probability of significance of the word is expressed by Zipf's law. Zipf's law allows to asses the relevance of the terms and identifies their patterns for the corpus [4]. There are some drawbacks to Zipf's law, the formulation of Zipf's law is ambiguous, from statistical perspectives, also it is not suitable for the big corpus. Three versions of Zipf's law are designed. The versions are tested on more than 30, 000 words. Statistical tests are used for fitting of functions. It's resulted in the fitting of more than 60 % terms at 0.05 significance level. [5].

III. Research Methodology

The first step in the proposed approach is data collection and corpus creation. Different type's poems and stories written by different authors were collected from various websites [22-24]. BaSa [1] is applied to identify common tokens present in both Hindi verses and proses. Similarly, Zipf's law is applied to extract tokens for Hindi language and Common token extracted by Zipf's law for proses and verses are identified. Similarly same process is repeated for Marathi corpus, that is, Basa and Zipf's law is applied on Marathi corpus having prose and verses and common tokens are extracted. Finally, compare the tokens retrieved by Zipf's law and Basa for both language corpora.

Library Udpipe present in "R" programming language is used to perform different NLP operations such as tokenization, tagging, lemmatization and so on. Udpipe is language-agnostic and can be trained given annotated data. Udmodel is a function of udpipe to load language type that is either Hindi or Marathi. Fig. 1 shows a diagrammatic representation of research methodology.

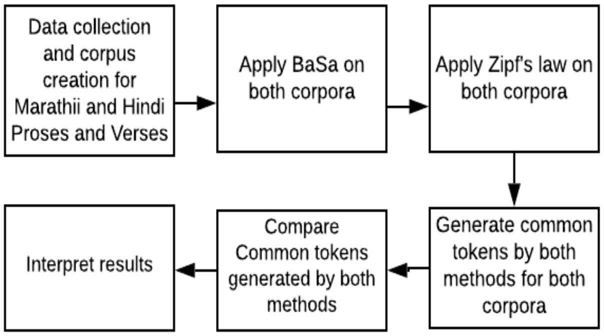

Data collection and corpus creation for Marathii and Hindi Proses and Verses
Apply BaSa on both corpora
Apply Zipf's law on both corpora
Generate common tokens by both methods for both corpora
Compare Common tokens generated by both methods
Interpret results

Fig 1. Diagrammatic Representation of Research Methodology.

Data collection and corpus creation: Different types of poem and stories written by different authors were collected from various websites. Hindi Corpus includes 820 stories and 710 poems and Marathi corpus includes 610 stories and 505 poems.

Implement BaSa on Marathi as well as Hindi corpus and identify common tokens.

BaSa includes tokenization followed by stop words and noise removal, normalization, TF-IDF and formation of synsets. Thus it involves context-based identification of common terms. It means that if the word ‘raat’ means night is present in the story and nisha is present in the poem , it identifies that these are synonyms and considers it as one synset group, thus in a final step comparison between synsets groups of verses and poems is being carried out. Similarly, synset groups of Marathi corpus are being constructed for e.g. ‘Aayusha’ and ‘Jeevan’ means life is being considered in a synset group and accordingly common tokens are identified.

Apply Zipf’s law to extract tokens on Hindi as well as Marathi corpus. Zipf’s law does not follow preprocessing and other NLP steps. It is based on the frequency and rank of the term. For eg. एकदा भक्त पुरंदरदास राजवाड्यात गेले होते, (Ēkadā bhakta purandaradāsa rājavāḍyāta gēlē hōt) (Once, the devotee, Purandaradas went to the palace). भक्ताने त्या तांदळात थोडे हिरे मिसळले होते. (Rājānē tyā tāndaḷāta thōḍē hirē misaḷalē hōtē) (The king had mixed little diamonds in the rice.). Zipf’s law will consider ‘होते’ as the maximum frequency word that is 2. Rest all words have frequency 1 and thus the rank of ‘होते ‘is 1. Thus Frequency and Rank are inversely proportional.

Identify common token extracted by Zipf’s law on the corpora (proses and verses) of Marathi and Hindi Zipf’s law is applied on the corpus of Marathi and tokens which are present in both verse and proses are found out.

Compare common tokens identified by Zipf’s law and BaSa on the corpora (proses and verses) of Marathi and Hindi: In the last step, common tokens identified by both of the methods are compared. It is observed that common tokens generated using BaSa are slightly more than tokens extracted by Zipf’s law.

IV. Results and Discussions

Table I presents the detailed token analysis used by BaSa. The first column shows the sample statement of a Marathi poem, by considering space as delimiter tokenization is achieved, that is separating all words, punctuations numbers and so on. The third column removes stopwords, “ते", "म्हणजे" and so on removed. In the next step of lemmatization, the word is converted into its root form that is “माणसाच्या" means man’s is reduced to "माणूस means man. The effect of TF-IDF can be reflected for multiple documents. Combining TF-IDF with synset construct the group of similar terms together and treated as one synset group.

TABLE I. Steps in BaSa Approach

Sample statement in the corpus of Marathi Poem	Tokenization	Removings topwords	Lemmatization	TF-IDF+Synset
माणसाच्या सुखाचं व आनंदी रहायचं एकमेव रहस्य ते म्हणजे हास्य,	"माणसाच्या",: सुखाचं", "व ","आनंदी" ,"रहायचं", "एकमेव", "रहस्य", "ते" ,"म्हणजे" ",हास्य", ","	माणसाच्या", :सुखाचं", "आनंदी""रहायचं", "एकमेव","रहस्य", ",हास्य"	"माणूस","सुख" "आनंद", "राहणे" , "एकमेव" , "रहस्य", ,हास्य"	**सुख, आनंद हास्य** ", "राहणे" , "एकमेव", "रहस्य"," माणूस"

Table II states the number of extracted tokens and the common tokens at each level of BaSa for Marathi as well as Hindi corpus. The corpus consists of 505 verses and 610 proses. At each stage, 50 to 60 % tokens are reduced. For both languages. NLP results are almost the same for both languages’ proses and verses and it proves that prose and verse behave same that is neutral as per as NLP activities are concerned.

Zipf’s law is used to decide important tokens of the corpus. It is executed on the corpus of Hindi stories/proses. It is based on the frequency and rank of the token. Table III shows the token frequency and its corresponding rank. It's clear that rank and frequency are inversely proportional. The last column specifies the significance level which shows importance of the term based on probability measure.

Rank Frequency graph allows to visualize word/token rank versus token frequencies based on Zipf's law. It is clear from the graph that the rank of the token is minimum for the highest frequent word and rank increases as the frequency decreases. The tokens are extracted from corpus of Hindi stories. Fig. 2 shows Rank-Frequency plot for the corpus of the Hindi proses Same way Zipf’s law is executed on Marathi corpus. The sample of tokens extracted from Marathi Poems is presented in the table. It is observed that the maximum word frequency of the term is 500.

TABLE II. Summary of sample corpus at each stage of BaSa for 505 poems and 610 Stories

Sr.No	Corpus	Verses		Proses		Common tokens	
		Hindi	Marathi	Hindi	Marathi	Hindi	Marathi
1	Total number of tokens	77,282	75,144	1,29,260	88,505	45,157	43,891
2	After removing stop words	35,181	32,234	54,239	45,321	22,123	19,234
3	Lemmatization	18,282	16,123	23,202	22.123	11,231	10,340
4	TF-IDF	7,139	7,139	8,102	7,123	2,123	1,123
5	Synset (groups)	4034	3,543	4,203	3.912	916	910

TABLE III. ZIPF'S LAW MEASUREMENT FOR THE CORPUS OF HINDI STORIES

Sr.no	word	Frequency	Rank	Significance level
1	घुस	4	1	0.6688
2	चिल्ला	3	2	0.5033
3	छाँह	2	3	0.4133
4	रिक्शे	1	4	0.3722
5	बच्चों	1	5	0.3613

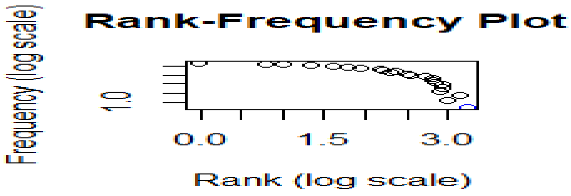


Fig 2. Rank Frequency plot by Zipf's law.

So according to Zipf's law, top and bottom 20 % tokens are discarded. So words having a frequency between 100 to 400 are considered. Eg. "तू","सगळ्यात", "तिला", won't be considered as major tokens and will be discarded. Table IV shows tokens with their frequencies with respect to the entire corpus.

Fig. 3 shows sample tokens extracted using Zipf's law for Marathi corpus. It can be observed that terms above 20 % threshold are being considered. X-axis shows the terms and Y axis shows the frequency of the term. The line shown in the graph is known as the Pareto line which is used to represent a cumulative percentage to show the importance of the term.

TABLE IV. TOKEN FREQUENCY FOR SELECTED TOKENS BY ZIPF'S LAW

Word	तू	सगळ्यात	आधी	आजी	बोलायचं	तिला
Frequency	500	459	410	389	210	90

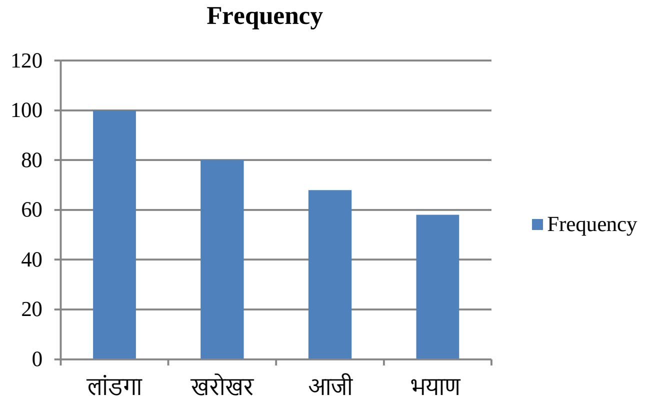


Fig 3. Common Tokens Identification by Zipf's law.

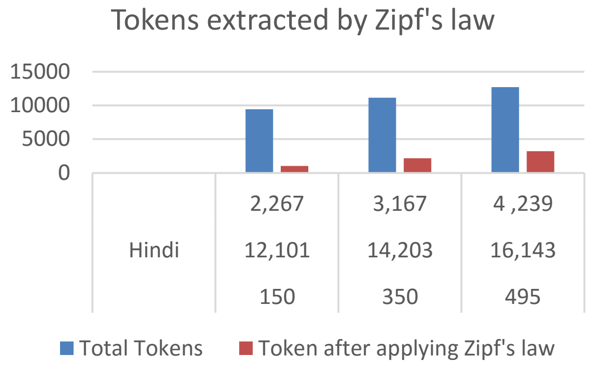


Fig 4. Token Count Extracted by Zipf's law.

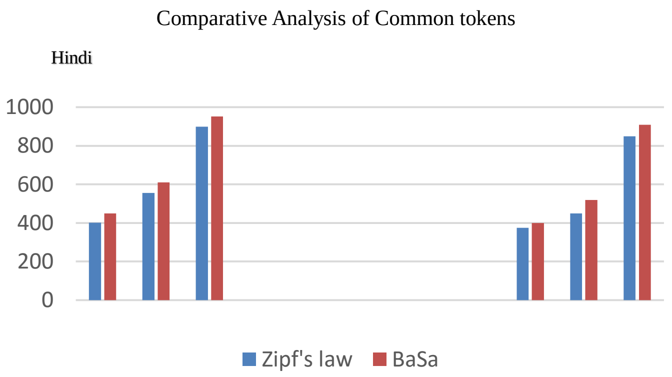


Fig 5. A Comparative Analysis of Common Tokens Extracted by Zipf's law and BaSa.

Fig. 4 shows a total number of tokens extracted and tokens retrieved after applying threshold using Zipf's law on Marathi and Hindi corpora for varied data size. There is more than 60% reduction in terms after applying Zipf's law. For the Hindi corpus size of 150 verses, total tokens are 12,101 and selected tokens are 2,267.

Fig. 5 shows a comparative analysis of the common tokens on varied corpus size of both the languages. X axis shows the sample data of verses and proses. Y axis shows common tokens. The number of common tokens extracted by BaSa is slightly more than Zipf's law. Not only the count of the token is more, but also the quality of tokens is better with respect to the context of the term. This is due to synset grouping varied out by BaSa.

V. CONCLUSIONS

Zipf's law is applied for corpora of proses and verses. NLP activities were carried out using BaSa. BaSa proved to be better than Zipf's law. Prose and verse give same results as per as NLP activities are concerned, so researchers can take either proses or verses or both for performing NLP activities. Hindi and Marathi corpora were considered and more than 3 Million documents were processed to identify common tokens between verses and proses. Considering Hinglish words (English words written in Hindi) can be incorporated in future.

References

[1] Bafna P.B., Saini J.R., 2020, BaSa: A Context based Technique to Identify Common Tokens for Hindi Verses and Proses, under review

[2] Mishra, D., Venugopalan, M., & Gupta, D. (2016). Context specific Lexicon for Hindi reviews. Procedia Computer Science, 93, 554-563

[3] Jena, M. K., & Mohanty, S. (2019, December). Predicting Sensitivity of Local News Articles from Odia Dailies. In International Conference on Biologically Inspired Techniques in Many-Criteria Decision Making (pp. 144-151). Springer, Cham]

[4] Wyllys, R. E. (1981). Empirical and theoretical bases of Zipf's law.

[5] Moreno-Sánchez, I., Font-Clos, F., & Corral, Á. (2016). Large-scale analysis of Zipf's law in English texts. PloS one, 11(1).

[6] Milan Straka and Jana Straková. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe(http://ufal.mff.cuni.cz/~straka/papers/2017-conll_udpipe.pdf). In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universaldependencies, Vancouver, Canada, August 2017

[7] Saini J.R. and Kaur J., 2020, "Kāvi: An Annotated Corpus of Punjabi Poetry with Emotion Detection Based on 'Navrasa'", Procedia Computer Science, in press with Elsevier

[8] Bafna P.B., Saini J.R.,2019, "Identification of Significant Challenges in the Sports Domain using Clustering and Feature Selection Techniques", 9th International Conference on Emerging Trends in Engineering and Technology on Signal and Information Processing (ICETET-SIP-19, Nagpur, India, in press with IEEE.

[9] Bafna P.B., Saini J.R.,2019, "Hindi Multi-document Word Cloud based Summarization through Unsupervised Learning", ", 9th International Conference on Emerging Trends in Engineering and Technology on Signal and Information Processing (ICETET-SIP-19), Nagpur, India, in press with IEEE.

[10] Bafna P.B., Saini J.R., 2019, "Scaled Document Clustering and Word Cloud based Summarization on Hindi Corpus, 4th International Conference on Advanced Computing and Intelligent Engineering, Bhubaneshwar, India, in press with Springer.

[11] Bafna P.B., Saini J.R., 2020, On Readability Metrics of Goal Statements of Universities and Brand-promoting Lexicons for Industries, 4th International Conference of Data Management, Analytics and Innovation (ICDMAI 2020), Delhi,India

[12] Bafna P.B., Saini J.R., 2020, Identification of Significant Challenges Faced by Tourism and Hospitality Industry Using Association rules", 4th International Conference of Data Management, Analytics and Innovation (ICDMAI 2020), Delhi,India

[13] Bafna P.B., Saini J.R., 2020,"Marathi Text Analysis using Unsupervised Learning and Word Cloud", International Journal of Engineering and Advanced Technology,9(3),in press

[14] Bafna P.B., Saini J.R., 2020, "Hindi Verse Class Predictor Using Eager Machine Learning Algorithms" International Conference On Emerging Smart Computing And Informatics 2020(IEEE-ESCI-2020), Pune, India.2018 (March)

[15] Bafna P.B., Saini J.R., 2020,"On Exhaustive Evaluation of Eager Machine Learning Algorithms for Classification of Hindi Verses", International Journal of Advanced Computer Science and Applications, in press

[16] Bafna P.B., Saini J.R., 2020, Hindi Verse Class Predictor using Concept Learning Algorithms, International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2020)

[17] Venugopal G., Saini J.R., Dhanya P., Novel Language Resources for Hindi: An Aesthetics Text Corpus and a Comprehensive Stop Lemma List, International Journal of Advanced Computer Science and Applications, vol. 11(1), Jan. 2020, in press

[18] Bafna, P., Pramod, D., & Vaidya, A. (2016, March). Document clustering: TF-IDF approach. In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (pp. 61-66). IEEE.

[19] Harrag, F., & El-Qawasmah, E. (2009, August). Neural Network for Arabic text classification. In 2009 Second International Conference on the Applications of Digital Information and Web Technologies (pp. 778-783). IEEE.

[20] Yu, B., Xu, Z. B., & Li, C. H. (2008). Latent semantic analysis for text categorization using neural network. Knowledge-Based Systems, 21(8), 900-904.

[21] Zhang, D., & Lee, W. S. (2003, July). Question classification using support vector machines. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 26-32).

[22] https://proudtobeprimary.com/reasons-teach-poetry-classroom/

[23] https://www.livehindustan.com/nandan/good-morning-story/news

[24] http://www.cfilt.iitb.ac.in

Logical Intervention in the Form of Work Breakdown Strategy using Object Constraint Language for e-Commerce Application

Shikha Singh[1], Dr.Manuj Darbari[2]
Department of Computer Science and Engineering
School of Engineering
Babu Banarasi Das University
Lucknow (Uttar Pradesh), INDIA

***Abstract*—This paper proposes a framework for Rule Based inhibition on e-commerce website for prevention of double payment and computing time invariant for concurrent event handling. Authors have analyzed computational models in terms of Customer segmentation replicating their buying characteristics and Dialogue level constraint establishment through OCL. The tool used are MDT-OCL and matching logic for logic level interpretation. The MDT tool generates Context Syntax Tree. We have used LPG grammar to be applied WBS codification by differentiating Descriptive Noun containing description about the products and associated verbs about the product. Authors have used Eclipse plug-ins to embed it logic constraint mapping to check for any ontological errors and double selection and payment errors.**

***Keywords*—OCL; e-commerce; concurrent handling; work breakdown structure; augmented querying**

I. Introduction

This Nowadays there is change in trend from physical movement of client to supplier's location and finally purchasing a product, its digital millennium e-commerce website has gained considerable importance. There are various categorization of e-commerce i.e.B2B, B2C & C&B. The basic idea is to provide ease of use to the client be it any category of client. In all the e-commerce applications main focus is on processing delay and high throughput in terms of outcome as desired by the user. Object constraint language was used in many earlier literatures but failed to achieve high throughput as processing time was significantly high. Chomicki [1] proposes a efficient method on semantic interconnections. Barbara [7] also suggested efficient method for component based modeling for e-commerce website.

All the above proposals have some drawback or the other; for example a client way want to be guided to right product, but unfortunately the time as well as the outcome of the query are main cause of concern. Proper merging from the various databases is a big problem.

This paper proposes a method to enhance the linking of various databases, by forming clusters thereby increasing the throughput of the query. Secondly as compared to normal queries Argumentative Query provides much increased output using OCL logic foundations. By embedding theoretical formal syntax providing an Argumentative logic to reach to a final efficient query outcome.

The paper starts with a brief introduction discussing about the current problems in e-commerce website, followed by exhaustive literature survey on OCL and its application in e-commerce. The third section proposed solution taking the case study of Amazon website. The proposed solution is supported by Analysis showing how augmentative constraint language is related with satisfaction level and response time. Lastly the paper is supported by conclusion and future scope of the work.

II. Literature Review

Author in [2] emphasis on use of fuzzy sets and possibility theory as a realistic approach for the representation of various categories of constraints. Author in [3] represents how transaction concepts apply to distributed systems and how to build high performance, high availability applications with finite budgets and risks. Author in [4] discusses a systematic approach towards preference engineering, preferences as soft constraints and a query model, along with its algebra and decomposition. Deductive arguments and counterarguments were been discussed with formalizing them using logical languages [5]. Author in [6] discussed the transformation of Unified Modeling Language (UML) models into software execution models and simulate the performance prediction for an e-commerce application. Author in [8] propose a web perusing history mining based client inclination revelation technique for web based business frameworks. Another strategy called UPSAWBH (User Preference Similarity Calculation Algorithm Based on Web Browsing History), which measure the degree of clients' inclination likeness based on their website page click designs, is advanced. Author in [9] presented, a preference update model to address the problem which arises, when customers issue requests based on out of date information in e-commerce application systems. They also proposed a group evaluation strategy for preference update processing in multi-database environment. Approaches for controlling the social actions that web 2.0 applications allow users to execute [10]. A method to enhance the accuracy of the model by adding static semantics to the SMP2 Meta model is proposed [11]. The control over these actions is defined with UML/Object Constraint Language (OCL) and then demonstrated through a prototype system. Verbalization of business rules deciphers the principles communicated in a design language into semi-regular articulations. This permits business specialists to approve models communicated in a plan language without suggesting any aptitude on this language [12]. A change instrument is proposed to robotize verbalization and applied to OCL (Object Constraint Language) limitations in the utility area. A structure to help formal displaying and agreements for information driven web administrations is show to be utilized as to confirm accuracy properties for synthesis of services [13]. OCLLib, OCLUnit and OCLDoc are proposed in [14]. OCL lip makes easier for the development of OCL expressions and constraints. It makes a high reuse factor which is configurable and testable named as OCL unit and OCL DOC. A security policy using Object Constraint Language has been developed in [15] secure MOVA tool utilized to answer the result of the proposed approval verification of non-trivial security properties. An access control Meta model has been developed [16] using unified modeling language (UML). This modeling language is independent of access control requirement using genetic mutation along with independentness from specific implementations. An experiments framework has been proposed in [17], for evaluation the constraints in object constraints language, along with Automated support for OCL refactoring. Modeling Business process is one of the important issues in maintaining competitions and dealing with challenges in business environment. Model driven software engineering is the new paradigm for the designing software. For meta modeling, UML has been used to model the problem along with a static semantic of the language specified by the object constraint language (OCL) as discussed in detail in [18]. In [19] temporal and history based authorization constraints are represented by the OCL. Also, first orders linear temporal logic has been utilized for formally satisfy the constraints. OCL is playing an important role in object oriented software development in the framework of UML and Eclipse modeling framework (ENF). In [20] an extension of OCL, Soil (Simple OCL-like Imperative Language) has been proposed, using declarative representation. Author in [21] presents a UML composition formalization that strictly adheres to the UML specification. This formalization paves the way for future work to check the consistency between a design model and its implementation in terms of the UML composition. Author in [22] presents a combined Object Constraint Language (OCL) and Object Role Model (ORM) for integrity constraints modeling, and demonstrates an implementation which enforces them by using a commercially available DBMS. Generally, OCL constraints are written manually, which may cause incorrectness and extra overhead. Therefore, generating OCL constraints template for UML models is a superior solution and is presented in [23]. The OCL constraints template automatically generated can be used as a reference for software designers. Author in [24] makes an investigation into the recent developments and explores the role of OCL in the current scenario and its future applications. Author in [25] presented an ontology-based approach for verification of business processes. They specified business rules as a logic program and used ontology reasoner for discovering model elements which violate the rules. Author in [26] exhibited ontology-based data intensive EIS (Enterprise Information System), which is notification- oriented. Author in [27] proposed domain specific language called ReSA for an embedded system. The ReSA utilizes axioms of ontology for specification of the embedded system. They perform scalable formal verification of various Simulink models. Sunitha, E. V. et al. 2018 [28], UML is achieved by using hundred percentage automation in code technology system fashions will make a drastic development in software program enterprise. This model examines a way to enhance the code age from UML models, with the assistance of “OCL”. It likewise investigates the potential outcomes to combine OCL in UML action models and produce code from the OCL unrivaled movement diagrams. Hammad, Muhammad et al. 2017 [29], The “OCL” is generally utilized intended for identifying an extra restriction on representation. To assist practitioners as well as researchers to indicate “OCL” restriction, they intended and advanced an Internet-based device referred to as interactive OCL (iOCL) for interactively identify a restriction on a given version. The middle concept behind iOCL is to here and show only appropriate information (e.G., operations) of “OCL” to customers at a specified pace of restriction requirement procedure, further toward assisting modelers with its syntax. Thus, they finish that iOCL can assist the system of “OCL” restriction requirement. Zaragoza, Mechelle Grace et al. 2018 [30], The method explains software of cellular integration additives in public organization and e-commerce as a software improvement technique to basically incorporate the specific primary mechanism of the era right into a solitary net-primarily based explanation. They tested a systematic improvement process for the software agent using additives and UML. They initially organized the agent factor arrangement and form it. Based on this, we evolved a cellular application for social commercial enterprise packages. They incorporate the module-based software program system into Drupal’s content control system. Author in [31] proposed the utilization of security presentation flexibility expectation working method.

III. Proposed Solution

This paper proposes a constraint language in e-commerce application with emphasis on cognitive search and Behavioral transposition of customers during buying process. Various computational models are identified and analyzed and is discussed in Table I.

TABLE I. A Comparison of Various Computational Models

	LEVELS			
Features	Dialectical	Logical	Dialogue	Rhetorical
Argument	Which Argument wins	Arguments are Atomic in Nature	Agents Exchange Arguments in Various Activities	Aim of Argument is wider
Methodology	Arguments & Counter Arguments	Arguments taken from Knowledgebase	Dialogue games are created in the form of Communicative Acts	By way of Persuasion using Threats and Rewards

In order to apply these computational model to check the buying behavior we first develop the logical foundation of Rules base as suggested by E. Franconic et al.

OCL: Set:: =OCL – Set ⟶ Union (OCL - set)
OCL - Set ⟶ elect (Var | OCL - Bool) |
OCL - Set ⟶ reject (Var | OCL - Bool) |
Class.allInstances

The main issue with the above problem is application of constraints in processing their query with large amount of concurrent request to be using the limited resource. This happens with the case of sale announced by various companies during peak and off peak season. Here comes the role of computational model using Argumentative models and work Breakdown strategy to provide better customer satisfaction and prevention of chocking at the server end. Consider an e-commerce site where a customer wants to buy a wrist watch and there is another query which is related to stainless steel watch strap only, then both queries will have the same resources to be shared we can apply Argumentative models for evaluating the customer's request (Fig. 1).

OCL logical Rule Base has been formed like in this case it can be written as:

Context wrist watch inv Watches:

self.purchase = 'any'implies.self.watch⟶ steelstrap.

Can be satisfied using a mix of preference model which will discriminate the customer based on the priority and its relationship with the resource availability i.e. by way of Argumentative Rules applied with OCL to form the Work breakdown structure (WBS).

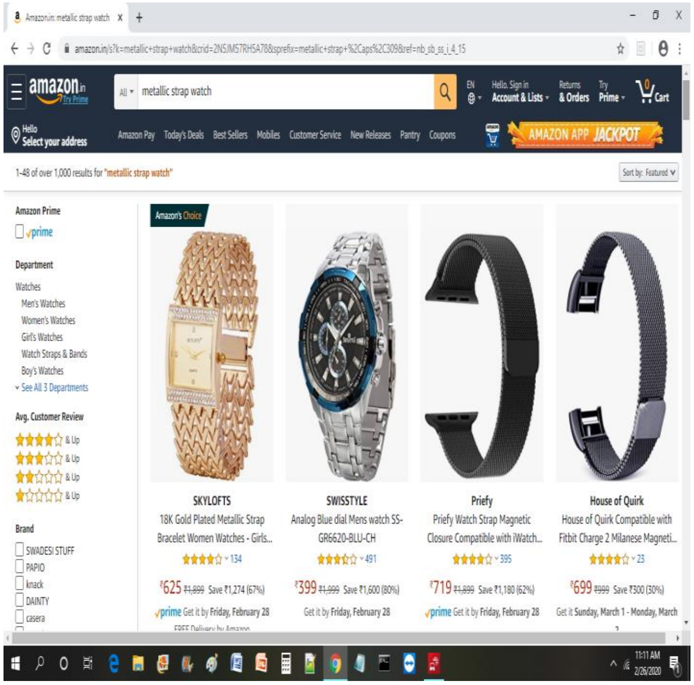

Amazon.in: metallic strap watch
amazon.in/s?k=metallic+strap+watch&crid=2NSJM57RH5A78&sprefix=metallic+strap+%2Caps%2C309&ref=nb_sb_ss_i_4_15
amazon.in
metallic strap watch
Hello, Sign in Account & Lists
Returns & Orders
Try Prime
Cart
Hello Select your address
Amazon Pay
Today's Deals
Best Sellers
Mobiles
Customer Service
New Releases
Pantry
Coupons
AMAZON APP JACKPOT
1-48 of over 1,000 results for "metallic strap watch"
Sort by: Featured
Amazon Prime
Department
Watches
Men's Watches
Women's Watches
Girl's Watches
Watch Straps & Bands
Boy's Watches
See All 3 Departments
Avg. Customer Review
Brand
SWADESI STUFF
PAPIO
knack
DAINTY
casera
Amazon's Choice
SKYLOFTS
18K Gold Plated Metallic Strap Bracelet Women Watches - Girls...
₹625 ₹1,899 Save ₹1,274 (67%)
Get it by Friday, February 28
SWISSTYLE
Analog Blue dial Mens watch SS-GR6620-BLU-CH
₹399 ₹1,999 Save ₹1,600 (80%)
Get it by Friday, February 28
Priefy
Priefy Watch Strap Magnetic Closure Compatible with iWatch...
₹719 ₹1,899 Save ₹1,180 (62%)
Get it by Friday, February 28
House of Quirk
House of Quirk Compatible with Fitbit Charge 2 Milanese Magneti...
₹699 ₹999 Save ₹300 (30%)
Get it Sunday, March 1 - Monday, March 2
11:11 AM
2/26/2020

Fig 1. Snapshot of wrist watches purchase.

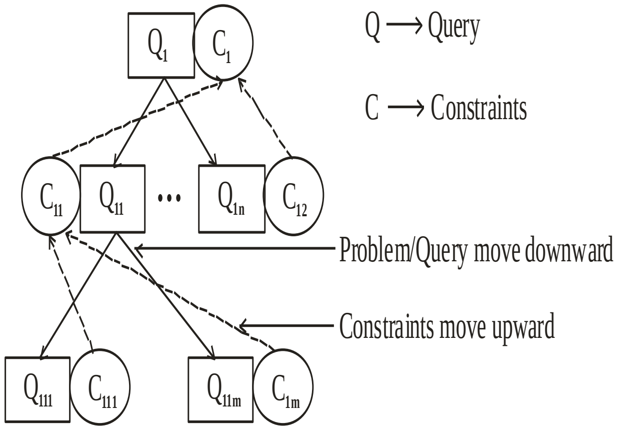


Fig 2. Task Reduction being achieved with Query and Constraints (OCL).

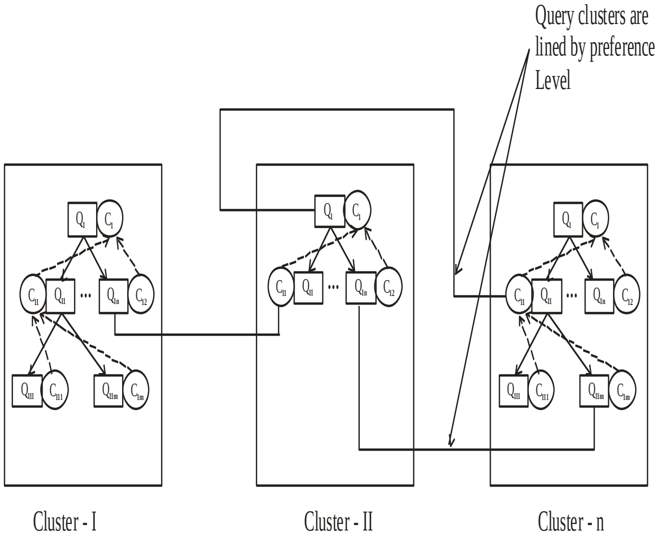


Fig 3. Cluster formation of Queries with OCL Logic.

The above Fig. 2 represents the Task Reduction technique as suggested by (Michanel &Bourdan, 2007). This is combined with Argument Centric preference Adaptation Model with OCL Logic constraints at each and every layer.

The basic idea of this model is to break down the searching item query by creating a cluster and then applying the cluster constraints to solve the problem.

These clusters in Fig. 3 represent the particular group with concurrent request of resources.

Let 'Q' be the instance which is a result of two instances Q1 and Q2 requesting the particular resources.

Now according to argumentativeness the constraint should be applied on the Model which can be written as

Q. allInstances ( )⟶ for All r | Q1.allInstances ( ) ⟶exists (cluster1 | cluster2| cluster n)

--------- Equation - I

A. The decision Making process using Argumentative Constraints

Set 'Q' be the query with 'S' as the solution derived in the form of equation-I, above, we write the rules using temporal logic in Object constraint logic as argumentative graph:

Q_1 = {S_1 OK (Constraint set 1), S_3 ^ (Constraint set 3) $\longrightarrow$ $S_1 \wedge \neg S_3$}

Q_2 = {S_4 OK (Constraint set 4), S_3 ^ (Constraint set 3) $\longrightarrow$ $S_4 \wedge \neg S_3$}

.

.

.

Q_n = {S_m OK (Constraint set m), S_{k-1} ^ (Constraint k-1) $\longrightarrow S_m \wedge \neg S_{k-1}$

Where:

Q1 =Watches

Q2 = Metallic Watches

Q3 = Metallic Strap Watches

Q4 = Watch with Metallic Dial

.

.

.

S1 = Best choice is Garmin

S2 = Best choice is Apple

S3 = Best choice is Samsung

S4 = Best choice is Fitbit

S5 = Garmin with Metallic Watch

S6 = Apple with Metallic Watch

S7 = Apple with Metallic Dial

The above constraint of Object Argumentative solution is derived by the help of dialogue rule clusters arranged in Work Breakdown manner where the Queries moves downward and constraint move upward to narrow down the search and also avoid concurrent processing.

Each cluster can be represented as presented in Fig. 4.

Layer Number	Flow of Query	Constraints Used	Solution Derived
1	{Q_1} →	{C_3} →	S_1
2	{Q_3} →	{C_2} →	S_3
3	{Q_4} →	{C_5} →	S_2
4	{Q_5} →	{C_4} →	S_4
5	{Q_2} →	{C_1} →	S_5

Fig 4. Block Inside representation of cluster.

We are able to derive the Flow of information with constraint to check two factors:

(i). One is Concurrency of the Request

(ii). Second is Ease of use to the user with Argumentative OCL at the backend

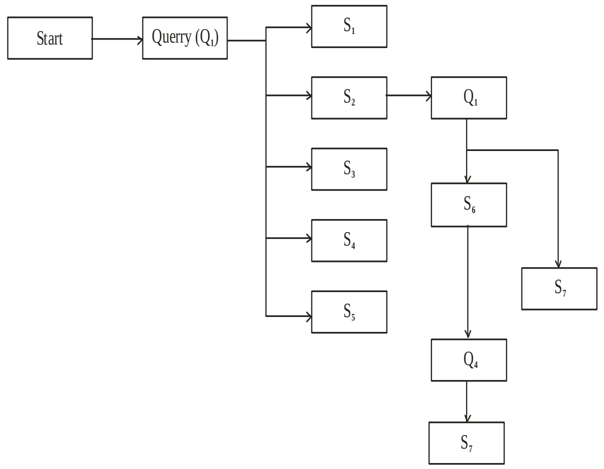


Fig 5. Argumentative OCL has been used to derive the solution S_7.

The above WBS protocol constraint shows the breakdown of Query set using Rhetorical Level constraint forcing the Client Query to leave S_6 Resource and move to S_7 solution, which is the derived solution set discussed in Fig. 5.

B. Analysis

Based on the Online purchase of Metallic watch, we represent the success rate of our model for standard query resulting into the desired output with a mean of three Query Request per hour we are to plot the success rate and Query processing time using Argumentative constraint Language as shown in Fig. 6 and Fig. 7, respectively.

The above two comparative figures shows the graph of satisfaction level and response time using MySQL normal Query and Argumentative OCL embedded with MySQL Query interface. Number of queries was generated and regards to the watch purchase and Argumentative Query Language showed much efficient output in terms of Response time and Satisfaction Level as the right output.

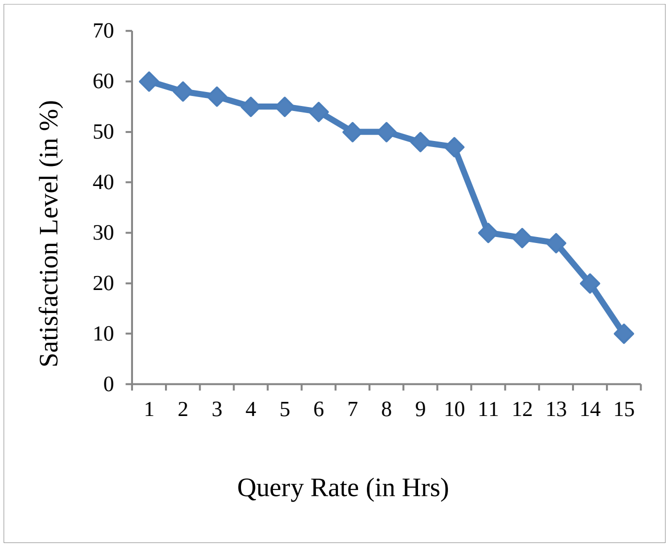


Fig 6. Success Rate of Argumentative Constraint Language.

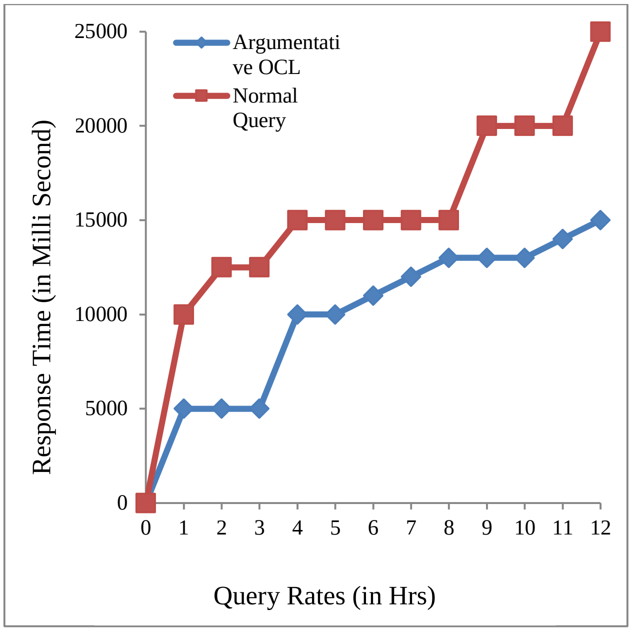


Fig 7. Comparison of Response Time.

IV. Conclusion and Future Scope

The paper discusses the issue of application of constraints on Query search for a particular product, and then the satisfaction level in terms of the right matches is analyzed. The paper proposes the up gradation of normal OCL by using Argumentative computational model thereby enhancing the capability of search in terms of more persuasion and directional.

The paper suggests how augmentative computational method provides a more formalized and refined method of querying. The problem of time of query is simplified and satisfaction level has significantly increased along with drop in response time with respect to normal query. It was achieved by using cluster in which various sets of work breakdown Ontologies are stored and are retrieved used argumentative centric preference adaption using OCL.

Future work relates to developing a parser where we can directly embedded the rules of argumentative search and it can retrieve from it.

Acknowledgments

My Sincere Thanks to my Guide and mentor Dr. Manuj Darbari Sir without whom this work would not have been possible.

References

[1] Chomicki, J. (2019) Preference formulas in relational queries, ACM Transaction of Database Systems.

[2] Dubois, D. (1996) Possibility theory in constraint satisfaction problems: Handling priority, preference and uncertainity. Applied Intelligence, 6, 287-309.

[3] Gray, J & Reuter, A.(1993) Transaction Processing : Concept and Techniques, Morgan Kaufmann, Sam Mateo, C.A.

[4] Kiessling, W. Foundatin of Preferences in Database Systems. Proc of 2002 IEEE Int. Conf. on Data Mining (ICDM '02).

[5] Cialdini, R : Influence : The Psychology of Persuasion, Harper Cellins (1984).

[6] Besnard, P., Hunter, A : Constructing Arguments graphs with deductive Arguments : A tutorial Argument & Computation 5(1), 5-30 (2014).

[7] Barbara and Massimo, Designing components for e-service, Proceedings of VLDB workshop on Technologies for e-services, Cairo, Egypt, 2000.

[8] D Evange Geetha, Ch Ram Mohan Reddy, T V Suresh Kumar, K Rajani Kanth, " Performance Modeling and Evaluation of e-commerce Systems Using UML 2.0", proc of 8th IEEE, ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (2007).

[9] P.Li, M.Tu, Z.Xia, "Preference update for e-commerce applications: Model, language, and processing", Springer Science + Business Media LLC 2007, Electronic Commerce Res(2007)7: 17-44.

[10] Z.Maamar, V.Buregio, N.Faci, D.Benslimane, Q.Z.Sheng, "Controlling " Web 2.0 applications in the workplace,19th International Enterprise Distributed Object Computing Conference(EDOC), pp.191-200, 2015.

[11] L.Haibing, Z.Ning ,L.Yonglin,L.Xiaobo ,Z.Yifan "EMF Based Validation Methods of the Static Semantics of Models", 2nd International conference on Information Science and Control engineering, IEEE, pp. 207-211, 2015.

[12] R.Baghli, B.Traverson ,"Verbalization of business Rules Application to OCL Constraints in the Utility Domain", 2nd International conference on Model Driven engineering and software development , IEEE, pp. 348-355, 2014.

[13] I.Saleh, G.Kulczycki, M.B.Blake, "Formal Specification and Verification of Data Centric Service Composition", International conference on Web Services (ICWS) IEEE, pp. 131-138, 2010.

[14] C.-O. Joanna, OCLLib, OCLUnit, OCLDoc: pragmatic extensions for the object constraints language, LNCS vol. 5795, 2009, 665-669.

[15] Basin, M. Clavel, J. Doser, M. Eaea, automated analysis of security design models intonation and software technology, vol. 51 no. 5, 2009,pp 815-831.

[16] S. Nadera, K. Hemanth, A. Kamel, L. Luegi, VACML: unified access control modeling language, 2011 4th IFIP international conference on new technologies, modeling and security (NTMS), Canada, pp.1-8.

[17] A.L. Correa, C.M.L. Werner, M. Barros, Refactoring to improve the under standability of specifications written in object constraint language, IET, software, vol. 2, pp-69-90.

[18] C. amelunxen, A. Schurr formalizing model transformation rules for UML/MOF2,IET software, vol. 2 no 3,(2008).pp. 204-222.

[19] K. Sohr, M. Drouineaud, G.J. ahn, M. Gogolla Analyzing and managing role based access control polices, IEEE transactions on knowledge and data engineering, vol. 20(2008),pp. 924-939.

[20] F. Buttner, M. Gogolla, Modular embedding of the object constraints language into a programming language, LNCS, vol.7021, 2011 pp. 124-139.

[21] H.M. Chavez, W. Shen , "Formalization of UML Composition in OCL", 11th IEEE/ACIS International Conference on Computer and Information Science. 2012.

[22] J. Saetent, N. Vejkanchana, S.Chittayasothorn, "A Thin Client Application Development using OCL and Conceptual Schema", 6th IEEE International Conference on Internet Technologies and Secured Transactions, Abu Dhabi, UAE , December 2011.

[23] Li Tan, Z. Yangand, J. Xie, "OCL Constraints Automatic Generation for UML Class Diagram", IEEE 2010

[24] Er. Shivani Gupta, Dr. Dhavleesh Rattan, "Research Paper on Implementation of OCL Constraints in JAVA" International Journal of Advanced Research in Computer Science, ISSN No. 0976-5697,Vol -8 , June-2011.

[25] Corea, C., and Delfmann, P., "Detecting Compliance with Business Rules in Ontology-Based Process Modeling", Proceedings of 13th Internationale Tagung Wirtschaftsinformatik, pp. 226-240, 2017.

[26] Liao, Y., Panetto, H., Simão, J.M., and Stadzisz, P.C., "Ontology-Based Model-Driven Patterns for Notification-Oriented Data-Intensive Enterprise Information Systems", Proceedings of 7th International Conference Information Social Technology, pp. 148-153, 2017.

[27] Mahmud, N., “Ontology-Based Analysis and Scalable Model Checking of Embedded Systems Models” M{ä}lardalen University, 2017

[28] Sunitha, E. V., and Philip Samuel, "Object constraint language for code generation from activity models", Information and Software Technology 103 (2018): 92-111.

[29] Hammad, Muhammad, Tao Yue, Shuai Wang, Shaukat Ali, and Jan F. Nygård. "iOCL: an interactive tool for specifying, validating and evaluating OCL constraints." The science of Computer Programming, Vol. 149 (2017), pp.3-8.

[30] Zaragoza, Mechelle Grace, Haeng-Kon Kim, and YounKy Chung. "Components of Mobile Integration in Social Business and E-commerce Application." In International Conference on Computational Science/Intelligence & Applied Informatics, pp. 59-68. Springer, Cham, 2018.

[31] Pathak, Nitish, B. M. Singh, and Girish Sharma. "UML 2.0 based framework for the development of secure web application." International Journal of Information Technology 9, no. 1 (2017): 101-109.

Adaptive Scheduling Design for Time Slotted Channel Hopping Enabled Mobile Adhoc Network

Sridhara S.B[1]
Professor, Dept. of ECE
Vijaya Vittala Institute of Technology
Bengaluru, India

Ramesha M[2]
Assistant Professor, Dept. of ECE
GITAM School of Technology,
GITAM (Deemed to be University)
Bengaluru, India

Veeresh Patil[3]
Professor, Dept. of ECE
Amruta Institute of Engineering and Management Science
Bengaluru, India

***Abstract*—Industrial Internet of things (IIOT) applications comprises the wearable sensor devices for human activity monitoring; these devices generate the continuous data at higher data rate and it is powered through the battery. Hence, it restricts the uses of wireless protocol such as IEEE 802.15.4 and BLE (Bluetooth Low Energy). Moreover, there are promising technologies such as TSCH (Time Slotted Channel Hoping) MAC (Medium Access Control) which can be deployed in the different environment, which are prone to interference. In this research work we focus on overcoming the issue for designing the AS (Adaptive Scheduling) for TSCH – MANET (Mobile Adhoc Network); furthermore it is very difficult to design scheduling technique considering the unpredictable nature of data source location and wireless link, this results in waste of reserve resource. Moreover proposed Adaptive Scheduling model allows the both slots i.e. shared and dedicated slots, it also allows the communicating device to active the assigned slots adaptively. Hence, our proposed AS model achieves the higher overall access fairness, minimal idle listening overhead, higher packet deliver rate; further to cope up with the higher traffic load MANET device can activate the additional slots dynamically. Moreover, the outcome of Adaptive Scheduling model shows the higher data transmission and lower energy consumption.**

Keywords*—6TiSCH; access fairness; energy efficiency; MANET (Mobile Ad Hoc Network); TSCH (Time Slotted Channel Hopping); scheduling

I. Introduction

Modern applications service of communication requires different characteristics such as scalability, adaptability, dependability and low idleness, However to develop such model it requires high maintenance and high cost. Furthermore, advancement in wireless tries to fulfill the performance guidelines through proposing new mechanism. IEEE 802.15.4 (2015 standards) is recently developed prototype and it is known as the modified version of IEEE802.15.4 (2006 standards) [1]. It acts as the empowering agent remote PAN (Personal Area Network), it possesses the characteristics of SCFR (Short Communication Frequency Range), low bandwidth and low powered [2], best example of this kind model is Sensor Network and MANET(Mobile Adhoc Network). 2015 standards of IEEE 802.15.4 develop the TSCH mode, which is very much feasible in providing the strength to fading ANF multipath interference. Furthermore rise of IPV6 over 6TiSCH [3] helps in providing the Internet Protocol organized abilities in case of low powered TSCH and thus it fills the gap between the modern technologies environments.

In 6TiSCH model, the data link layer is available on the IEEE 802.15.4-TSCH, in 802.15.4 TSCH MAC, MANET adopts the scheduling technique and it synchronizes on the frame organization. However, 802.15.4 – TSCH fails to explain modelling of scheduling method. In recent years several researchers has been developed for packet scheduling in the 6TiSCH, furthermore it is observed that scheduling is modelled through distributed or centralized manner. However In earlier method, MCS (Main Control Server) is responsible for the constructive scheduling mechanism, Moreover n case of decentralized environment condition, global system information and main controller in formation is not accessible. Hence, MANET devices acknowledge the scheduling mechanism through the neighboring devices by process of schedule optimization and interchange change of data in device to device. Centralized technique [4][5] are mainly suitable for the static network where overhead is nominal, further the centralized based scheduling mechanism is suitable for the large MANET device, however the cost plays an important role. In case of decentralized technique, this is mainly applicable mainly for the huge scale system and which is dynamic in its nature. Hence it is observed that in comparison with centralized, decentralized based brings the better energy utilization.

Moreover, energy efficient results in reduction in energy overhead [6] for MANET routing [7]. IoT observes the huge application such as smart cities, smart homes and wearable devices, further the IoT application requires flexible and robust characteristics, and apart from this minimum energy, consumption is required and further support the huge operation. In this research work, the main aim is to enhance the energy efficient and data transmission of TSCH network. Moreover, TSCH is parted based on high data rate and dynamic traffic; moreover, this is characterized due to dynamic nature of WNC (Wireless Network Channel) condition. Hence, retransmission amount is random, further In case of High Bandwidth Network the packet reception capability of the MANET Device is utilized to its full potential. Nevertheless, this phenomena occurs mainly due to the various aspects, first aspect is that it is highly improbable to forecast the data rate and mobile location, second aspect is that it is very difficult to adopt and deal with dynamic nature of the wireless network, third it is very difficult to find the change in route. Meanwhile it is also observed that it timescale variation in these changes are in terms of milliseconds. In the above, three condition the scheduling does not fit, this means it is non-practical to design and share in reactive fashion. Hence, to avoid that static allocation is needed over the slot allocation, however this might result in reduction in completely scheduling capability and this might result in many reserve slots of TSCH left unused. Moreover, it is also observed that conservative over the allocation might cause in packet loss increment, this occurs due to lack in scheduling capacity catering the high volume traffic is difficult. Similarly, there might occur energy overhead due to excessive TSCH slot allocation. Furthermore, the assigned slot are not fully utilized by the transmitter and causes the idle listening overhead in receiver. Hence, to overcome the above-discussed problem this research work proposes the AS aka Adaptive Scheduling for the MANET (TSCH enabled). Adaptive Scheduling aka AS is designed through extending Static scheduling this further allows the receiver and transmitter to choose the time-slots based on the properties of slot utilization.

The contribution of work is as follows:

- This paper presents the novel AS aka Adaptive Scheduling model for mobile Adhoc Network that is TSCH enabled.
- Adaptive Scheduling achieves the satisfactory tradeoff between the energy and packet transmission performance.
- The main advantage of Adaptive Scheduling is that it provisions for the high traffic application and this does not have any effect on the energy consumption of MANET, furthermore this is achieved through activating the additional slots dynamically and it does not require the reorganizing of new schedule.
- The Adaptive model outperforms various state-of-art technique such as [8], [10], [11], [12], [13] and [14] in terms of access fairness, packet transmission and energy efficiency.

The paper is articulated as follows: Section I gives brief introduction of scheduling mechanism using TSCH for wireless and mobile adhoc network. Further, highlights research problem, issues and challenges in designing adaptive scheduling design. In Section, II the proposed adaptive scheduling model for TSCH enabled mobile adhoc network is presented. Experiment result and analysis is discussed in Section III. Lastly, the conclusion with future research direction of work is discussed.

II. Adaptive Scheduling Model for TSCH Enabled Mobile ADHOC Network

In this section, we discuss the AS aka Adaptive Scheduling for TSCH enabled MANET, further this particular research focus on designing the efficient scheduling mechanism, which achieves the satisfactory tradeoff between the maximizing packet transmission and minimizing the energy efficiency. Here at first system model is design, followed by that we design energy model for computing the energy dissipation per packet, later the tradeoff model is developed for energy minimization, last but not least this section presents the Adaptive Scheduling.

A. System Model

Let's consider the high traffic network that has high data rate and it comprises the huge amount of MANET devices that performs under the dense mobile adhoc network, however considering such environment TSCH based slotted aloha does not perform well due to collision occurrence. Meanwhile packet transmission is restricted once the slots are allotted in individual manner and it causes restriction in catering devices. Moreover let us consider the TSCH network (single Hop based), which comprises the dingle sink device and two-end device X and Y, these end devices communicates with the bandwidth s_X and s_Y respectively. Further, these end devices towards the sink are describes through the DLL(Data Link Layer) packet reception rate q_X and q_Y respectively, meanwhile this research work considers the retransmission as the independent Bernoulli distribution Set that has the similar PRR in case of each distribution. Moreover, equation 1 computes the cumulative predictable transmission denoted as S_X for the device X.

$$S_X = s_X \sum_{o=1}^{\infty} o q_X (1-q_X)^{o-1} = \frac{s_X}{q_X} \tag{1}$$

In similar fashion, for device Y predictable number of transmission is computed and denoted by S_Y

$$S_Y = s_Y \sum_{o=1}^{\infty} o q_Y (1-q_Y)^{o-1} = \frac{s_Y}{q_Y} \tag{2}$$

S_X in addition, S_Y are known PPS (Packet per Slot)-frame, amount of slots in the particular slot frame is denoted byO_G, where $O_T \leq O_G$ indicates the slots shared within the slot frame; furthermore here each MANET device comprises $\frac{1}{2}(O_G - O_T)$ contention less slots.

Here each end devices tries to utilize the dedicated slot first i.e. Until O_E packets for each slot frame, later it transmits the residual packet, hence excess PL (Packet Load) is formulated through the below equation.

$$D_X = S_X - O_E \tag{3}$$

And

$$D_Y = S_Y - O_E. \tag{4}$$

In case of non-availability of shared slots$(O_T = 0)$, excess packet loss is also assumed lost; else if the slots are available then it is communicated through shared medium with collision $\frac{D_X}{O_T}$ and $\frac{D_Y}{O_T}$ in respective manner. Later cumulative packet collision which is denoted by L is formulated through the below equation and considered as the likelihood of end devices X and Y that chooses the same slots.

$$L = O_T \frac{D_X}{O_T} \frac{D_Y}{O_T} = \frac{D_X D_Y}{O_T}. \tag{5}$$

Further end-to-end PDR (Packet Delivery Rate) is computed foe end device X through the below equation.

$$PDR_X = \begin{cases} 1 - \frac{D_X}{S_X} & O_T = 0 \\ 1 - \frac{L}{S_X} & O_T > 0\ and D_X \leq O_T \\ 1 - \frac{L+D_X-O_T}{S_X} & O_T > 0\ and D_X > O_T. \end{cases} \quad (6)$$

Similarly PDR (Packet Delivery Rate) for the end device Y is computed through the below equation.

$$PDR_Y = \begin{cases} 1 - \frac{D_Y}{S_Y} & O_T = 0 \\ 1 - \frac{L}{S_Y} & O_T > 0\ and D_Y \leq O_T \\ 1 - \frac{L+D_Y-O_T}{S_Y} & O_T > 0\ and D_Y > O_T. \end{cases} \quad (7)$$

Furthermore, the system performance is optimized through the shared timeslot O_T and maximize the average PDR (Packet Delivery Rate).

B. *Energy Consumption Model*

In order to design the energy consumption model this particular research consider the contention free schedules, hence at first unicast transmission is assumed and here each slot time is categorized into TxRx slot, idle listening slot and sleeping slot. Furthermore, transmitter and receiver is considered to be in the sleeping state and their radio is turned off. Moreover TCH scheduled slot is considered as the unused slots, further cumulated energy dissipation is computed through the below equation (this is considered in the sleeping stage):

$$E_{slp} = O_{slp} \cdot \left(2. O_{slp}. W\right), \quad (8)$$

In the above equation, O_{slp} is electric charge in the sleeping slot process for the given voltage W, here W indicates the MANET supply voltage. Moreover, factor 2 indicates both receiver as well as transmitter, in case of idle listening slot; for initializing the communication between the transmitter and receiver a particular slot is dedicated. However In here none pending packet is observed among the sender, Hence energy wastage is induced in case of Idle listening slot, further the cumulated ED(Energy Dissipation) can be computed through the below equation:

$$E_{idlsn} = O_{idlsn} \cdot \left(U_{slp} \cdot W + U_{idlsn} \cdot W\right), \quad (9)$$

Moreover in above equation, U_{idlsn} indicates induced electric charge among the receiver for idle listening slot stage; further In case of TxRx slot , for communication between the transmitter and receiver, a particular slot has been dedicated. Hence, Transmitter is switched on to perform on the transmitter and meanwhile the receiver observes and listens to the given channel for further communications. Meanwhile, pending packet queues possesses minimum one frame for performing the transmission, at last cumulated ED (energy Dissipation) for the TxRx slot can be computed through the below equation:

$$E_{trnsrecv} = O_{trnsrecv} \cdot (U_{TrnsDatRecvAck} \cdot W + U_{RecvDatTrnsvAck} \cdot W), \quad (10)$$

Moreover in above equation, $U_{TrnsDatRecvAck}$ indicates the induced electric charge to perform the data transmission among the transmitter and acknowledgement packet, similarly $U_{RecvDatTrnsvAck}$ depicts the induced electric charge among receiver. Nevertheless, these parameters are described as the upper limit in case of failed transmission; this occurs due to channel error or any presence of interference. Moreover, in this work, it is taken as the successful transmission and further it is incremented with $O_{trnsrecv}$. Hence, consumption of average energy per packet is formulated as below.

$$\mathbb{E} = \frac{E_{slp}+E_{trnsrecv}+E_{idl}}{s \cdot FRAMES} \quad (11)$$

Here s indicates the packet load incoming, which is further described as the packets per frame, similarly $FRAMES$ indicates the total number of simulated frames.

C. *Energy Minimization Tradeoff Model*

In this sub-section, we focus on the trade-off model between the reliable packet transmission and energy minimization for the TSCH enabled MANET; further, it is assumed that there are eight active slots always active for transmission. Later we minimize the active slots while reducing energy dissipation of radio. In case if the slot size is still less than eight then this might not be sufficient for the relaying the packet load, in case if slot size is still more than 8 then due to idle listening the energy gets wasted. Hence, to achieve the good trade-off we define μ is the energy per packet to perform the packet delivery efficiently [14], [15] [16].

$$\mu = \frac{\mathbb{E}}{PDR^{o}}, \quad (12)$$

In the above equation o helps in regulating the tradeoff specification of given reliability over the energy dissipation, meanwhile lower value of μ indicates better scheduling of MANET (TSCH based) [17], [18], [19].

Moreover proposed Adaptive scheduling model provide assurance of access fairness among the adjacent contending devices, further in the communication slot, given MANET device can select any adjacent device which has non-empty data packet buffers as receiver. Furthermore it is observed that employing the adjacent device for the personal gain leads to the bandwidth starvation in case of low prioritized devices, mainly due to the condition where the data in adjacent queues are not transmitted before the expiration. In this work, the selection process has been carried out in the round robin manner through utilizing the available timeslots as input; further considering O as the MANET device, which has the information regarding device's channel offset [20]. Moreover, the proposed algorithm provides the equal priority to the neighbor until the slot frame length and O are of co-prime numbers, this provides the proposed adaptive scheduling to minimize the packet drop rate in comparison with the existing model. Hence, considering the above scenario, it is clear that proposed model assures the access fairness; later it minimizes the energy consumption of given MANET devices and similarly we achieve the better packet transmission in comparison with the existing model. Moreover, the proposed model is evaluated in the next section of this research work.

III. Result and Discussion

In this section, the Adaptive Scheduling is evaluated and compared with the existing scheduling technique [8], further for evaluation we have used I-5 processor packed with 12 GB RAM; 6TiSCH simulator is used and it is written using the python programming language through the associate of 6TiSCH WG [9] which is open source. Moreover, the existing model [8] and proposed model are incorporated in given 6TiSCH simulator; furthermore, the simulation parameter is given in Table I and parameter consideration is in accordance with the industrial environmental condition where the heavy traffic load is occurred [10]. Evaluation is carried out in terms of packet routing performance and energy overhead through comparing existing and propose AS model.

A. *Energy Consumption Perforamcne Evaluation*

This particular section gives the comparison analysis of proposed model over the existing model in terms of energy consumption; the evaluation is performed by considering the various transformation rate and packets. Moreover, Fig. 1. depicts the comparison of existing and proposed model by varying packet number; in here, the packet variation is from 3600 to 7200. Furthermore, through the Fig. 1, it is observed that energy consumption of proposed Adaptive Scheduling for 3600, 4800, 6000, 7200 packets is 15.22%, 2283%, 27.99% and 32.78% lesser than the existing model. Meanwhile the average energy consumption is reduced by 24.77% when compared to the existing model.

B. *Throughput and Packet Delivery Rate Perforamcne Evaluation*

In this section, the comparative analysis of existing and proposed model based on packet routing performance such as throughput is evaluated, the evaluation is carried out varying the transmission rate(in Mbps). Fig. 2 shows the throughput comparison of existing and proposed model, here we observe that performance of conventional static algorithm is satisfactory till the TR (Transmission rate) of 6 MBPs. In case of higher transmission rate the existing model underperforms whereas proposed Adaptive Scheduling performs satisfactory until 20 Mbps, furthermore proposed model achieves the 47.83% enhancement in comparison with the existing model.

C. *Access Fairness Performance Perforamcne Evaluation*

In this sub section of performance evaluation the comparative analysis of existing and proposed model is carried out based in the access fairness; further the evaluation is carried out by varying the number of packets. Fig. 2 shows the packet drop rate comparison of existing and proposed model; it is observed that packet drop rate of proposed model is very low when compared to the existing model. Moreover, the packet drop rate of proposed model is reduced by 99.7% over the existing model, this shows the significant access fairness of proposed Adaptive scheduling model.

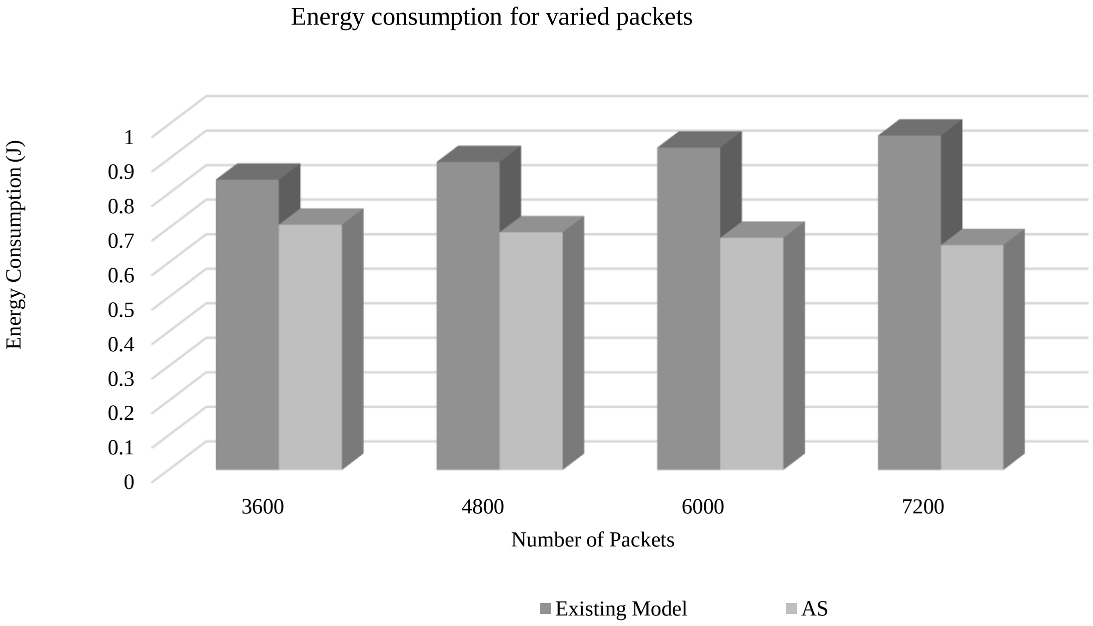


Fig 1. Energy Consumption Performance Evaluation Considering Varied Packets.

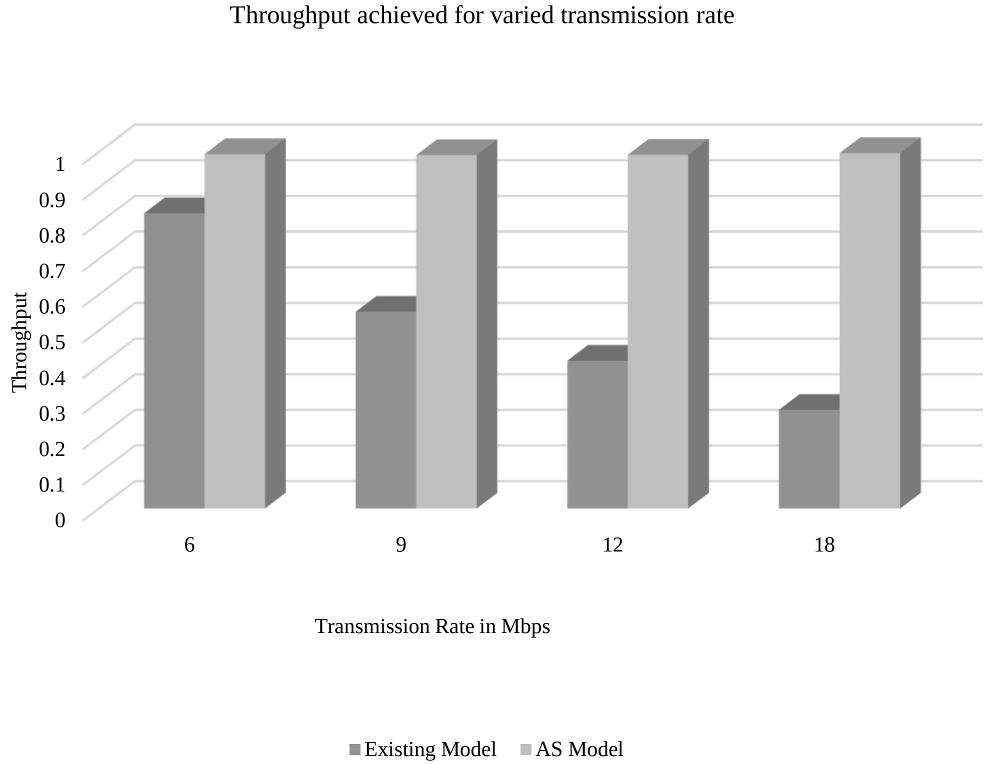


Fig 2. Throughput Performance Achieved Considering Varied Transmission Rates.

D. Comparision with Existing Model

In this section we discuss the evaluation of different model when compared to our exiting model; in [11] author carried out the huge survey of existing model by addressing the performance issues for the real time application. Moreover, [8] proposed a scheduling technique, which considers the decentralized network, and this technique achieves the 44.5% improvisation in throughput, however the energy consumption and packet routing performance is ignored. Similarly, [10] and [12] performed the evaluation and achieves the energy consumption minimization of 27.08% and 25.6%, respectively; whereas [13] achieved the energy overhead reduction of 22.38% and packet dropt rate of 25% when compared to the existing technique. Moreover, we compare our model with [8] since they considered the decentralized network; it is observed that AS technique achieves the satisfactory trade-off between energy overhead and routing performance. Hence from discussion it is clear that our model outperforms the various state-of-art technique [8], [10], [11], [12], [13], and [14] considering the carious parameter such as access fairness, packet delivery rate and energy overhead.

IV. Conclusion

In this research work, we consider an adaptive scheduling model for TSCH enabled MANET; through the extensive analysis, it is observed that static scheduling technique possesses the higher energy overhead; this is due to the heavy resource allocation. Moreover, it is quite difficult to develop efficient scheduling technique since the data source location and nature of wireless link are unpredictable. Hence, to achieve the good tradeoff a model named Adaptive Scheduling is proposed which allow both slots i.e. shared slots as well as dedicated slots; further, it also allows the communicating device to activate their assigned slots and this results in access fairness. Moreover to cope up with the high load traffic, MANET device can easily activate assigned slots; further the comparative analysis is carried to evaluate the proposed AS model over the existing one. Proposed AS model achieves the energy overhead minimization of 24.77% compared to the existing model; furthermore considering the throughput performance AS achieves 47.83% better and packet drop rate by 99.7% better than the existing model. Hence proposed Adaptive Scheduling technique achieves the higher access fairness than any of the state-of-art technique, further in future work we would be focusing on evaluating considering different performance parameter and under multi-hop TSCH network.

References

[1] IEEE Standard for Low-Rate Wireless Networks, IEEE Standard 802.15.4-2015, pp. 1–709, 2016.

[2] "IEEE Standard for Local and metropolitan area networks--Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs)," IEEE Std 802.15.4-2011 (Revision of IEEE Std 802.15.4-2006), pp. 1-314, 2011.

[3] D. Dujovne, T. Watteyne, X. Vilajosana, and P. Thubert, “6TiSCH: Deterministic IP-enabled industrial Internet (of Things),” IEEE Commun. Mag., vol. 52, no. 12, pp. 36–41, Dec. 2014.

[4] Z. Shelby, K. Hartke, and C. Bormann, The Constrained Application Protocol (CoAP), IETF RFC 7252, 2014.

[5] I. Juc, O. Alphand, R. Guizzetti, M. Favre and A. Duda, "Energy consumption and performance of IEEE 802.15.4e TSCH and DSME," 2016 IEEE Wireless Communications and Networking Conference, Doha, pp. 1-7, 2016.

[6] S. Bandyopadhyay and E. J. Coyle, "Minimizing communication costs in hierarchically-clustered networks of wireless sensors," Elsvier Journal of Computer Networks, vol. 44, no. 1, pp. 1-16, 2004.

[7] L. Karim and N. Nasser, "Energy efficient and fault tolerant routing protocol for mobile sensor network," in IEEE International Conference on Communications (ICC), Japan, pp. 1-5, 2011.

[8] Municio, Esteban & Latré, Steven “Decentralized broadcast-based scheduling for dense multi-hop TSCH networks”, Proceedings of the Workshop on Mobility in the Evolving Internet Architecture, Pages 19-24, 2016.

[9] T. Watteyne, K. Muraoka, N. Accettura, and X. Vilajosana. The 6tisch simulator. https://bitbucket.org/6tisch/simulator/src, 2015.

[10] Kralevska, Katina & Vergados, Dimitrios & Jiang, Yuming & Michalas, Angelos. (2017). A Load Balancing Algorithm for Resource Allocation in IEEE 802.15.4e Networks, 2017.

[11] Rodrigo Teles Hermeto, Antoine Gallais, Fabrice Theoleyre “Scheduling for IEEE802.15.4-TSCH and slow channel hopping MAC in low power industrial wireless networks” Journal Computer Communications archive Volume 114 Issue C, Pages 84-105, 2017.

[12] Thang Phan Duy, Thanh Dinh, and Younghan Kim “A rapid joining scheme based on fuzzy logic for highly dynamic IEEE 802.15.4e time-slotted channel hopping networks”, International Journal of Distributed Sensor Networks, https://doi.org/10.1177/1550147716659424, 2016.

[13] X. Fafoutis, A. Elsts, G. Oikonomou, R. Piechocki and I. Craddock, "Adaptive static scheduling in IEEE 802.15.4 TSCH networks," 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), Singapore, pp. 263-268, 2018.

[14] A. Elsts, X. Fafoutis, J. Pope, G. Oikonomou, R. Piechocki and I. Craddock, "Scheduling High-Rate Unpredictable Traffic in IEEE 802.15.4 TSCH Networks," 2017 13th International Conference on Distributed Computing in Sensor Systems (DCOSS), Ottawa, ON, 2017, pp. 3-10, 2017.

[15] Hermeto, Rodrigo & Gallais, Antoine & Theoleyre, Fabrice. (2017). Scheduling for IEEE802.15.4-TSCH and slow channel hopping MAC in low power industrial wireless networks: A survey.ComputerCommunications.114.10.1016/j.comcom.2017.

[16] Du, Peng & Roussos, George. (2012). Adaptive time slotted channel hopping for wireless sensor networks. 10.1109/CEEC.2012.6375374.

[17] M. Ojo, S. Giordano, An efficient centralized scheduling algorithm in ieee 802.15.4e tsch networks, in: Conference on Standards for Communications and Networking (CSCN), IEEE, 2016.

[18] Hammoudi, Sarra & Harous, S. & Aliouat, Zibouda & Louail, Lemia. (2018). Time slotted channel hopping with collision avoidance. International Journal of Ad Hoc and Ubiquitous Computing. 29. 85. 10.1504/IJAHUC.2018.094400.

[19] Diab, Rana & Chalhoub, Gérard & Misson, Michel. (2013). Overview on Multi-Channel Communications in Wireless Sensor Networks. Network Protocols and Algorithms. 5. 112. 10.5296/npa.v5i3.3811.

[20] Wu, Yafeng & Stankovic, John & He, Tian & Lin, Shan. (2008). Realistic and Efficient Multi-Channel Communications in Wireless Sensor Networks. Proceedings - IEEE INFOCOM. 12. 1193 - 1201. 10.1109/INFOCOM.2008.175.

JeddahDashboard (JDB): Visualization of Open Government Data in Kingdom of Saudi Arabia

Mashael Khayyat[1]
University of Jeddah, College of Computer science and Engineering
Department of Information Systems and Technology
Jeddah, Saudi Arabia

Abstract—**Open data is data that anyone can freely use, access and redistribute without financial, legal or even technical restrictions. Accordingly, all governmental and non-governmental organizations may publish data that they own open for various purposes on the Internet without any restrictions such as (climate statistics, education statistics, transportation, industry, water abstraction, etc.). Further, Open Government Data (OGD) initiatives are proliferated in every country including Kingdom of Saudi Arabia (KSA). OGD should supposedly escalate the transparency, collaboration, and participation of citizens towards using OGD. However, the presentation of OGD format may not be attractive enough to users and vice-versa the data may not be easy for them to understand and interpret. These stumbling blocks may dampen the use of OGD among citizens. The problems can be resolved through visualization of the available data sets and to represent these data in accordance to user preference. This research emphasizes on visualization efforts of OGD in KSA named JeddahDashboard (JDB) website. The aim of creating JeddahDashBoard is to visualize the published government data in KSA. The idea was inspired by the DublinDashBoard in Ireland where data and real-time information, time series index data, and interactive maps on vast aspects of the city are provided mostly in an interactive ways and attractive charts that are easy to understand. In order to create JeddahDashBoard, two tools were used "the tableau" and then "chart.js" because the later was simple and flexible. Finally, this paper shares researchers experience and challenges in establishing JDB.**

Keywords—Open Data; Open Government Data; visualization; Dashboard; Saudi Arabia; KSA

I. Introduction

In the nutshells, Open Data can be defined as "an approach to managing data so that it enables the structured free flow of non-sensitive information to those who have the need or interest in using this information. It allows different types of users to access, organize and use data in ways that make sense to them" [1], [2]. Or simply can mean the kind of data which is open for anyone to access, modify, reuse, and share. Open Data derives its label from various open-source, open government, open science, etc. [3]. Governments, independent organizations, and agencies have come forward to create more open data for free and easy access [4].

Open data is important because the world has increasingly adapted data-driven [5], [6]. But if there are restrictions on the access and the format of how those Open Data are presented, the idea of data-driven business and governance will be difficult to be realized. Therefore, the quality of open data must be good which will enhance fuller understanding among citizens and can strengthen democracy. The understanding of Open Data is normally associated with the most impactful way the Open Data are presented, bearing in mind the purpose of sharing them with the targeted end-users. This directly demands data visualization perspectives. However, trying to decide which visualization type works for the specific Open Data, can be tough [7]. Users see various types of data visualizations each day. Some are beautiful but provide little insight. Some are functional, allowing the viewers to draw conclusions at a glance, but not aesthetic. Above all, the provider's goal, the structure, and size of the underlying Open Data will normally determine when to use one type of visualization over another.

Open data is not very new idea in the Kingdom of Saudi Arabia [8],[9],[10], there is already a website called www.data.gov.sa, which provides Open Government Data (OGD) to serve citizens, but the method of providing data could be not attractive to the users and the data can be difficult to be understood. Thus, to solve these problems we come up with the idea of visualizing the available data and the concept of our project is to represent this data in an attractive way so that more users can be interested to deal with the OGD and make use of it.

The next section explains the research background and related work. Then, experiment and results followed by discussions. Finally, conclusions and future work are described.

II. Research Background and Related Work

A. OGD and Principles of OGD Initiatives

Recently, government organizations and agencies have been adopting OGD initiatives globally to achieve many benefits such as institutional, social, cultural, economic and political benefits [11]. There are two major reasons for opening government data. Firstly, is the positive impact it will have on citizens which includes greater awareness of what the government does, cognizance of how their taxes are spent, and improved civic engagement. Next, the benefits that governments can realize such as increased civic trust in government, greater efficiency, and enhanced delivery of services or systems functions.

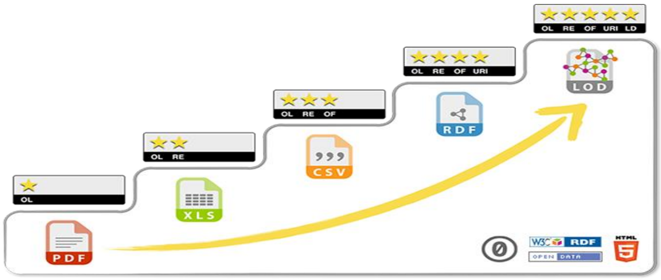


Fig 1. Classification of open data [17].

OGD has also been adopted among government institutions in KSA. As in [12], “The adoption of OGD was influenced by existing institutional arrangements and landscapes occurring in the country such as the Saudi Vision 2030, the approval of the Freedom of information Act 2016, and anti-corruption campaigns which have contributed positively to the transition from culture of secrecy to openness” [12]. In [12] further emphasized that the OGD initiative is influenced by both the internal and external institutional pressures. In [12] also confirmed that, the organizations involved in his study has obtained many benefits, which can be described as “rationalized myths” which he elaborate as “transparency and accountability, better access to government data, support for innovation, improved government services, operational benefits and encouragement of participation” [12].

The organization for Economic Co-operation and Development (OECD) [13] defines open government as the opening of government processes, proceedings, documents, and data for public scrutiny and involvement. Eight principles have been identified to guide open government initiatives. The principles [14], [15] include the following:

- All public data should be made available. "Public" data refers to information that isn't subject to valid privacy, security, or privilege limitations.
- Data is collected at its primary source, and it isn't modified or presented in aggregate.
- Data is made available in a timely fashion so that it's valuable and useful.
- Data is accessible to the widest number of users for the widest range of purposes.
- Data is structured so that it can be processed by a machine.
- Data is available to anyone, and no one needs to register to access it.
- Data is available in a non-proprietary format - no one has exclusive control over it.
- Data is license-free, and not subject to any copyright, patent, trademark, or trade secret regulations. However, reasonable privacy, security, and privilege restrictions are acceptable.
- Seven additional principles are added by the OpenGovData.org [16] to extend the OGD’s principles. The extension of these principles is more to improve the quality, the safety of the retrieval, the authenticity, and the integrity of the information published on OGD as specified below:

- Data should be free, and it should be available online.
- Data should be made available at a stable internet location for an indefinite period, and it should remain in a stable data format for as long as possible.
- Data should be trustworthy. To that end, it should be digitally signed or include an attestation of the publication/creation date, its authenticity, and its integrity.
- There must be a presumption of openness. That is to say, the government must be proactive about making information public and available.
- The government must provide users with enough information for them to determine whether the information is accurate and current.
- Data must be safe to open, without executable content that can transmit worms, viruses, and malware.
- The government implements suggestions from the public about how to disseminate information.

Keeping in mind that open data can be available but with deferent level or classification as Tim Berners-Lee proposed. He proposed a five-step ranking system for data on the web as show the Fig. 1 [17].

In Fig. 1, one star means make your data available online in any format under an open license. Two starts data means they are structured such as in Excel file format. However, if the data in CSV instead of Excel (a non-proprietary open format) they will considered as three stars. Four stars data means that URIs have been used in order to enable people to point at datasets. Finally, five stars means the data are linked with other data to provide context.

B. Open Data Visualization

Citizens' interests, technical know-how and purpose for using OGD varies. Visualization is necessary to facilitate citizens who are less competence in technical know-how to use, understand and/or contribute to OGD initiatives. There are different types of visualizations of OGD which include table, graph, chart, and map. There exist many tools that can be used to create visualizations of OGD. Among them are given in Table I.

C. OGD Problems and Proposed Solutions

Government Data is open if and only if, it can be accessed and reused by any Internet users. However, there may be barriers to open data [18]. The barriers can be from the perspectives of either financial, legal and technical issues or the combination of them. For example, the barrier from the financial constraints is when data is not free; from the legal barriers is when legal permission to access data is imposed and from the technical barriers is when data is only confined to certain formats such as PDFs and Microsoft excel worksheet files. These formats can be complicated to understand and use without visualization. Thus, this research aims to offer visualization types according to users' preferences. This initiative is hopefully facilitating citizens to access and use OGD while inspiring some of them to publish data when they realize what they share is leveraged by others.

D. Related OGD Initiatives Around the Globe

For the purpose of lesson learnt, five related OGDs around the globe are selected randomly using convenience sampling and summarized in Table II. It briefly describes the initiatives in the perspectives of names, objectives and techniques used to develop them.

TABLE I. VISUALIZATION TOOLS

Tool Name	**Brief description Table Column Head**
Chart.js	Chart.js is a community maintained open-source library that helps you easily visualize data using JavaScript. The library has two different versions. The normal version, called Chart.js and Chart.min.js, comes with the Chart.js library and a color parser. The bundled version composed of Chart.bundle.js or Chart.bundle.min.js.
Tableau	Tableau offers strong features for data discovery and features such as data conversion, sorting and filtering data. Tableau is a software suite that enables users to create data visualization that may be maps or graphs as it provides flexibility in creating different types of graphs. The files are CSV and databases rational and others. Tableau is also a reporting tool for findings of data analysis via a bar chart, pie chart and any visualization technique. Tableau has many extensions such as public addition, personal addition and professional addition. And the public addition connects many data sources such as: Google sheets, Microsoft excel 2007, Web data connectors and another data source.
Google Fusion Tables	Google Fusion Table is a tool to visualize data on the Excel file format as it allows users to create their own spreadsheets by uploading CSV files and allows charts from the spreadsheets that users have. However, Fusion Tables, the API and Embedded Fusion Tables visualizations (maps, charts, tables and cards) have been discontinued since August 2019 [19].
FusionCharts	FusionCharts is a javaScript charts for web & mobile. It's the most comprehensive JavaScript charting library, with over 90+ charts and 1000+ maps. From the basic charts (line, column, pie etc. 2D & 3D) to the most complex ones (waterfall, gnat, candlestick, zoom line etc.). With FusionCharts, it is easy to download/export all your JavaScript charts to the format of your choice - JPEG, PNG, PDF or SVG. All one need to do is to include a single line of code. FusionCharts Suite XT supports 3 modes of export namely, Server-side export, Client-side export, and Auto export (Default) (FusionCharts, 2018) [20].

TABLE II. RELATED OGD INITIATIVES

OGD Initiatives	**Objective**	**Technique**
OGD in Brazil	Brazil is one of the leading OGD providers. The objectives of Brazil OGD: ● To provide data without financial costs. ● To focus on data of economy, trade, environmental services, population and others. ● To encourage citizen participation in the planning and development of public policies.	Use many techniques and programming languages to implement the website which include: ● Database. ● SQL. ● Web programing. ● Power bi.
UK Government Data	Government of the United Kingdom has made data available to the public free of charge via the UK's open data website, data.gov.uk. Through this website, all governmental data issued by ministries and official institutions in UK. Open data are presented in several format, which include: PDF, JSON, HTML&RDF format. ● To provide data that meet the needs of consumers of such data. ● To provide easy data access. ● To provide good visualization varieties.	Use many techniques and programming languages to implement the website as expected: ● HTML. ● JAVA SCRIPIT. ● CSS. ● PHP. ● XML. ● Socrata tools.
The Dublin Dashboard (DDB)	The DDB Initiative is web based OGD system. DDB interacts with user by collecting, analyzing data from many sources about Dublin in Ireland through interactive maps, graphs and applications. DDB objectives are: ● To provides data for free. ● To provides fast and easy data retrieval to users, for example checking their traffic summons with interactive maps.	Use many techniques and programming languages to implement the website which include: ● HTML. ● JAVA SCRIPIT. ● CSS. ● PHP. ● XML ● Web programing. ● Project Open Data Dashboard tools.
Open Data portal of Saudi Arabia	The main role of the portal is to publish data sets for ministries and government agencies in the form of open data, and to make this data available to all users. ● To enable users to access and copy databases of different ministries and government agencies in Saudi Arabia and copy them and use them according to certain rules. ● To bridge the gap between government agencies and citizens, where citizens benefit from data provided in many ways.	Use many techniques and programming languages to implement the website which include: ● PHP. ● CSS. ● HTML. ● Web programing.

OGD Initiatives	Objective	Technique
	● To expand e-government services so that these efforts reach individuals and private sector organizations, ● To improve transparency and allow people to showcase their creations.	
Open data Princess Nourah bint Abdulrahman University	Princess Nourah bint Abdulrahman University offers an open data platform where all students and beneficiaries can benefit from the published data. The library offers an open data library on the University's website, which contains university files that include many statistics. The initiative has many objectives, including: ● To encourage electronic participation of governmental and non-governmental entities. ● To enable users to access data easily. ● To enhance transparency. ● To provides an open portal containing files of different data.	Use many techniques and programming languages to implement the website which include: ● HTML. ● JAVA SCRIPIT. ● Using software of Microsoft office such as Excel. ● Can files be viewed directly without the need to use specialized software.

III. Experiment and Results

A. Design, Development and Implementation of JDB

In this research, the gateway for sharing and publishing KSA's OGD is created and named "JeddahDashBoard" (JDB). JDB is incorporating visualization preferences from the perspective of users or citizens. JDB is inspired by the "DublinDashBoard", in Ireland, where real-time information, time-series index data, and interactive maps on all aspects of the city information are provided. On the same note, JDB specifically enables citizens to obtain detailed, up-to-date information about Jeddah city that will help them make a day-to-day decision and promotes evidence-based analysis.

The design, development and implementation of JDB is following the standard procedures of system development life cycle. JDB users' requirements were conducted via online questionnaires towards 200 respondents. The researchers get 75% feedbacks (150 of respondents returned the feedbacks). In summary, the findings showed the highest percentage of respondent is those in the age range of 21-26 years. Most of them is interested in public services data. Their motivation to use and deal with open data is to learn new skills. Highest percentage of data format preferred by the respondents is in the PDF format. The functional requirements should include starting up the website, displaying the homepage, exhibiting: The Browse Categories, the Display Visualized Dataset and the Read About Website. The non-functional requirements of JDB include concern of user friendliness of its website, its security, accuracies, accessibilities and availabilities.

The software used during design and implementation as well as managing JDB varies. The Operating Systems used is Windows (8 & 10) while the software for managing the JDB Project is Microsoft Office Project 2013. The Microsoft Office Word (2013) was deployed for the documentation purposes. The Microsoft Office PowerPoint (2013) and Prezi was used for presentation, while Google Drive and Microsoft Excel Project (2010-2013) were utilized to prepare for online survey and statistics. Microsoft Visio (2013) was deployed for drawing diagrams and Low Fidelity Prototype by mock-ups and Microsoft access (2013-2010) were arranged to design the prototype of JDB. Finally, PHP, Cake PHP, MySQL, Notepad ++, HTML, CSS, JavaScript, WAMP Server, Tableau, and Google Fusion were installed for JDB implementation.

There are nine database tables composed in JDB. The tables' names and their brief descriptions are listed in Table III.

100 participants were participated in the testing process. The breakdown of the participant is: 46 administrators (people in Municipality of Jeddah Governorate) to test and verify the administration functionalities of the JDB website and 54 users/citizens surround Jeddah (teachers and students of King of Abdul-Aziz University) to test the usability and reliability of the JDB website. The test was conducted from Sunday 25th March 2018 to Wednesday 28th March 2018.

B. Difficulties of JDB Initiative

As mentioned earlier, open data is not a very new idea in Saudi Arabia, and there is already a website called www.data.gov.sa, but the data are in spreadsheet format. JDB website offers visualizations to enable a user to view data according to the available charts she/he prefers which can facilitate his/her understanding of the data. For this study, researchers focus on open data from the education category first and make the data available only in three forms of chart namely: The Bar, the Pie and Line chart. Further, JDB is offering a useful function where a user can upload any excel file and convert the excel file to a chart of her/his preference (either a: bar, pie or line chart).

TABLE III. JDB Database Files Composition

Table Name	Description
About us page section	This table used to store page information
Contact us massage	This table contains messages from users who want to communicate with us.
News Letter	This table contains the email of users who subscribe to the latest site news.
Site settings	This table stores all figures and links of social media.
Admin	This table contains the admin who controls the website and updated.
Chart type	This table contains the chart types (Bar chart/Pie chart/Line chart).
File	This table contains the files in the website in each education data sets.
File chart	This table contains the file with selecting chart.

IV. Conclusion and Future Work

The aim of this work was to utilize available OGD by visualizing the data. Furthermore, this work aimed at representing OGD in an attractive way so that more users can be interested to deal with the OGD and make use of it. However, as mentioned above, there were many difficulties faced to utilize the available open data. The main challenge was identifying the appropriate tool/ technique to visualizing the OGD. As mentioned above, “chart.js” tool was used for enabling the users to upload any excel file to visualize the data. Again, the implementation stage was quite challenging where the researchers have encountered problems in identifying the appropriate tools to visualize the spreadsheet files. The researchers have changed the tools of visualization twice. In the beginning, they learned “the tableau” tool and how to link it to JDB website. Then, many modifications have been requested from the academic staff that made them change the tool to a simple but flexible Java Script charting named “chart.js”. The tool enables a user to upload any excel file to visualize the data.

Overall, there are many avenues to extend JDB initiative which can improve its quality and functionalities for the future research. The following include some interests:

- Extend the system to include all sectors (for example, Health, Taxation and Transportation) from the Saudi open data website (www.open.data.sa)
- Implement JBD system for mobile application
- Enable users to directly pull data from the "http://www.data.gov.sa/en" for live updated figures.

Further experimental studies are required for a detailed benchmark of this work with others. For example, JDB can be benchmarked with DublinDashBoard in Ireland where data and real-time information, time-series index data, and interactive maps on vast aspects of the city are provided mostly in interactive ways and attractive charts that are easy to understand.

Acknowledgment

I would like to thank prof. saadiah yahya for all her efforts and goodness, and my students malak al_nufaie, nourah al_subaie, salma al_muzaini, and shahad arbaein for all their contribution and efforts to achieve this work.

References

[1] Cortada, J. W., Nix, V. A., & Reyes, L. C. Opening up government: How to unleash the power of information for new economic growth. IBM Institute for Business Value. USA. 2011.

[2] ALGEMILI, Usamah A. Outstanding challenges in recent open government data initiatives. International Journal of e-Education, e-Business, e-Management and e-Learning, 2016, 6.2: 91.

[3] Khayyat, Mashael Mahmoud. CO-CREATION WITH OPEN GOVERNMENT DATA. 2017.

[4] Khayyat, Mashael, and Frank Bannister. "Open data licensing: More than meets the eye." Information Polity 20.4 (2015): 231-252.

[5] MÜLLER, Dominic; REICHERT, Manfred; HERBST, Joachim. Flexibility of data-driven process structures. In: International Conference on Business Process Management. Springer, Berlin, Heidelberg, 2006. p. 181-192.

[6] KADLEC, Petr; GRBIĆ, Ratko; GABRYS, Bogdan. Review of adaptation mechanisms for data-driven soft sensors. Computers & chemical engineering, 2011, 35.1: 1-24.

[7] WANG, Lidong; WANG, Guanghui; ALEXANDER, Cheryl Ann. Big data and visualization: methods, challenges and technology progress. Digital Technologies, 2015, 1.1: 33-38.

[8] SAXENA, Stuti. National open data frames across Japan, The Netherlands and Saudi Arabia: role of culture. foresight, 2018.

[9] ALRUSHAID, Marwah W.; SAUDAGAR, Abdul Khader Jilani. Measuring the data openness for the open data in Saudi Arabia e-Government: A case study. International journal of advanced computer science and applications, 2016, 7.12: 113-122.

[10] ALANAZI, Jazem Mutared; CHATFIELD, Akemi. Sharing government-owned data with the public: a cross-country analysis of open data practice in the Middle East. 2012.

[11] WANG, Hui-Ju; LO, Jin. Adoption of open government data among government agencies. Government Information Quarterly, 2016, 33.1: 80-88.

[12] Altayar MS. Motivations for open data adoption: An institutional theory perspective. Government Information Quarterly. 2018 Oct 1;35(4):633-43.

[13] OECD. 2018, [online] Available at: http://www.oecd.org/mena/governance/mena-oecd-open-government.htm [Accessed 2 Feb. 2020].

[14] Heusser, F. .Understanding Open Government Data and Addressing Its Impact: 2012.

[15] Sunlight Foundation. Ten Principles for Opening Up Government Information. [online] Available at: https://sunlightfoundation.com/policy/documents/ten-open-data-principles/, 2010.

[16] Opengovernmentdata.Org. Welcome to Open Government Data. [online] Available at: http://opengovernmentdata.org/ ,2016.

[17] 5stardata.info. 2017. 5-star Open Data. [online] Available at: http://5stardata.info/en/ [Accessed 2 Nov. 2017].

[18] Martin, S., Foulonneau, M., Turki, S., & Ihadjadene, M. Risk Analysis to Overcome Barriers to Open Data. Electronic Journal of e-Government, 11(1), 348-359. 2013.

[19] FAQ: Google Fusion Table [online] last update: 3 Dec 2019. Available at: https://support.google.com/fusiontables/answer [Accessed 17th Jan 2020] [

[20] Fusion Chart Development Centre [online] Available at: https://www.fusioncharts.com/dev/ [Accessed 17th Jan 2020].

Author’s Profile

Mashael Khayyat is now the supervisor of the department of Computer and Network Engineering, and an assistance professor in the Department of Information Systems and Technology in the College of Computer Science and Engineering at the University of Jeddah. M. Khayyat has graduated with a Bachelor of Computer Science with an honor degree in 2004 from King Abdul-Aziz University. M. Khayyat earned a Master of Applied Information systems (AIS) from the Arab Academy for Science and Technology and Maritime Transport, Alexandria, Egypt. M. Khayyat received her second Master's degree in Technology Management (MTM) from the University of New South Wales (UNSW), Sydney, Australia. In 2017, M. Khayyat has granted a Ph.D. degree in Computer Science and statistics from Trinity College Dublin (TCD), Dublin, Ireland.

Optimizing Genetic Algorithm Performance for Effective Traffic Lights Control using Balancing Technique (GABT)

Mahmoud Zaki Iskandarani
Faculty of Engineering
Al-Ahliyya Amman University, Amman, Jordan

***Abstract*—Genetic Algorithm (GA) is implemented and simulation tested for the purpose of adaptable traffic lights management at four roads-intersection. The employed GA uses hybrid Boltzmann Selection (BS) and Roulette Wheel Selection techniques (BS-RWS). Selection Pressure (SP) and Population (Pop) parameters are used to tune and balance the designed GA to obtain optimized and correct control of passing vehicles. A very successful implementation of such parameters resulted in obtaining minimum number of Iterations (IRN) for a wide spectrum of SP and Pop. The algorithm is mathematically modeled and analyzed and a proof is obtained regarding the condition for balanced GA. Such Balanced GA is most useful in traffic management for an optimized Intelligent Transportation Systems, as it requires minimum iterations for convergence with faster dynamic controlling time.**

Keywords—Genetic algorithm; traffic lights; intelligent transportation systems; correlation; roulette wheel selection; boltzmann selection; selection pressure; population

I. Introduction

As a result of population growth, increased level of pollution, people migration to urban areas from rural and to cities from urban areas, traffic congestion issue became more critical and pressing. Hence, Traffic lights optimization using adaptive and intelligent algorithms in conjunction with smart sensing is a must under these circumstances. This comes under design optimization, where smart algorithms such as genetic and fuzzy logic algorithms are employed to improve an existing transportation network as a function of increasing level of traffic flow levels, resulting in traffic congestion, delays, higher fuel consumption, air pollution and increase probability of traffic incidents [1-2].

In conjunction with population growth, a marked increase in the automotive industry is witnessed, where millions of vehicles are put on an existing roads. Monitoring the vehicular activities is an important issue to the transport authorities. Such large number of vehicles have a major impact on the environment and daily living of people using these roads due to congestion and delays [3-4].

The congestion issue forms a bigger problem on lanes and roads that are not structured appropriately, where it has a major effect on vehicular movements in a city. In particular at the intersections or with every region of traffic signal, especially during the peak hours. Shortening the duration of the traffic light signal, would not greatly improve traffic flow as important parameters, such as, que length and vehicle speed are not taken into account in a standard control algorithm [5].

Traffic congestion is a critical issue in affecting the lives of the society. Many areas suffered from long term socioeconomic damage owing to growing traffic congestion. To resolve urban congestion, conventional alternatives of increasing road capacity through road network expansion is limited in effectively reducing congestion, as large capital investment and stakeholders support is needed.

Intelligent Transport System (ITS) is used for the last few years in an effort to enhance efficiency and effectiveness of the existing transportation infrastructure by employing various sensor and communication technologies. Adaptive traffic signal control systems try to use the principles provided by artificial intelligence in an attempt to reduce congestion and provide safer traffic operations for both vehicles and pedestrians.

II. Background

Excessive Traffic density on the roads is a critical issue, as it leads to congestion. This is caused by rise in the number of vehicles and due to expansion and urbanization. Limitations on development and building of new highways and roads, initiated the need to optimize the use of existing infrastructure to achieve optimal flow of traffic. In addition, important time wasted because of traffic congestion, will implicitly affect productivity and performance, and thus affects people’s lives, both economically and socially.

Traffic light signal management and control has a marked impact on the efficiency and effectiveness of urban transportation systems. Conventional Traffic light signals are pre-programmed, pre-timed signals. Pre-timed control comprises a series of fixed duration intervals that are repeated continuously. Advanced Traffic light signals can operate in two ways:

1) Actuated mode: Actuated Traffic Light Signals detect and respond to the presence of vehicles or pedestrians at the intersection. They are supported by detectors within the intersection and the necessary control functionality to respond to traffic density and demand, in order to affect the signal cycles times dynamically.

2) Adaptive mode: Adaptive Traffic Light Signals control system continuously calculates optimal signal timings based on detected volumes and dynamically implement them. The system smartly and efficiently responds to the abrupt and fast changes in dynamic traffic conditions. It uses data from vehicle detectors so that traffic signal optimization is achieved.

Transportation network optimization (TNO) applied to traffic signal control, is implemented while considering route choice pattern of network users. It also include minimization of travel time and avoidance of new road construction. Conventional traffic lights operates using a constant switching cycle regardless of traffic load. Such static mechanism will not allow for variation of traffic load, events, emergency conditions, or general road incidents. Thus, there is a need for smart and adaptive algorithm that dynamically control traffic signals not only locally, but also all over the network for both vehicles and pedestrians. Such a critical solutions will carry out functional synchronization between traffic signals, in order to achieve a measurable reduction in congestion and delay levels, in addition to less pollution and safer driver and pedestrian roads. Such an objective can be achieved using genetic algorithm (GA) [6-10].

Genetic algorithms are used to imitate the processes of natural selection, where the best individuals have more probability to survive and their genes will be part of the creation of one or more offspring. Such process is repeated with the output of each new offspring is more fitted to survive than its parents. In the last few years, genetic algorithms found to present a suitable approach to complex transportation problems, as they are considered search algorithms that operate on the principles of natural selection. They determine a number of potential solutions within a population, with an encoding process that result in an optimized solution. In transportation, genetic algorithms are used to optimize and adapt the green interval response as a function of traffic density based on vehicles count [11-15].

Selecting the best parameters of a genetic algorithm, so as to obtain good results to optimize its performance, is very important to its effectiveness. Crossover, mutation rate and population size are the most influencing control parameters as reported by previous works [16-20]. However, pressure selection and population size in correlation is a new approach in balancing and GA algorithm optimization.

In this paper an investigation is carried out regarding selection pressure and population parameters within a GA algorithm used for intelligent traffic lights control and management, in order to obtain best adaptable performance under dynamically changing traffic density. The two parameters and their correlative effect on the performance of the employed GA algorithm is mathematically analyzed, simulated and results discussed.

III. Methodology

The main objective of this work is to optimize the designed genetic algorithm in order to produce an optimum traffic lights control mechanism, which is adaptable in nature and intelligent in behavior.

Selection process is a critical part in genetic algorithms, whereby chromosome is chosen from the available generation's population to be included in the next generation. Fitness function based process selects the best chromosomes, which is used to improve chances of individual survival.

The employed probability of selection process is based on two known algorithms:

1) Boltzmann Selection (BS): The algorithm is based on simulated annealing. Annealing is based on cooling such that a low energy state is reached. During the process, heating until melting is reached at a high temperature, through which, random movement is realized. The temperature is slowly cooled until the minimum energy states achieved. The equivalence between the optimization process in this work and the Boltzmann Simulated Annealing (SA) is the need to achieve stable levels with better new solutions (optimized traffic signal timing). Boltzmann Selection (BS), general expression is presented in equation (1).

$$P_{Selection}\left(_{i+1}\right) = \beta \exp\left(-\frac{y_{i+1} - y_i}{T_{i+1}}\right). \qquad (1)$$

Where;

y_i: Fitness function value for initial solution

$y_{i+1:}$ Fitness function value for the new solution

T: New artificial temperature

β: Normalization factor

In the algorithm used in this work, the rate of selection is managed by a continuously changing Delay Time (DT) parameter, which is equivalent to temperature in Boltzmann algorithm (simulated annealing). Initially Delay Time (DT) is high. So, one DT (Shortest Delay Time (SDT)) parameter will be at first high and decreases, with another DT (Longest Delay Time (LDT)) providing Delay Time as a function of increase in the testing population. (SDT) decreases gradually which increases the effect of Selection Pressure (SP). This results in determining and mapping the search space. In Boltzmann selection, the probability of selecting best value is high with lower execution time.

2) Roulette Wheel Selection (RWS): In this algorithm, selection of the fittest is carried out. The initial part of the selection process based on stochastic selection from one population to create the basis for the next population. In this process, the fittest have a better chance (probability) of survival than weakest ones. Thus, the fittest will move forward to the mating region to prepare for the next population. The process is shown in Fig. 1.

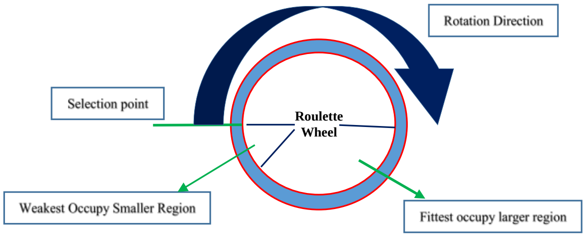


Fig. 1. Roulette Wheel Selection.

The number of times the roulette wheel runs is proportional to the size of the desired population. Whenever the wheel stops fittest individual will have a good chance of being selected for the next population and subsequent mating region.

The algorithm applies the expression in equation (2):

$$P_{Selection}\left(_{m}\right)=\left(\frac{y_m}{\sum_{k=1}^{n} y_k}\right) \quad (2)$$

Where;

n: Population Size

m: Number of times the wheel rotated

The combined Boltzmann Selection-Roulette-Wheel Selection (BS-RWS), for genetic algorithms (GAs) is based on both entropy and importance sampling methods. It naturally leads to adaptive fitness in which the fitness function does not stay fixed but dynamically varies.

Two important controlling parameters are used in optimizing and validating the used genetic algorithm for a four-traffic lights intersection covering four roads (Rd.1, Rd.2, Rd.3, Rd.4), each having a capacity of 60 vehicles:

1) Selection Pressure (SP)
2) Population (Pop)

Selection Probability (SProb) for traffic lights control, which in effect manages the number of vehicles passing from road to road at the four-road junction is governed by equation (3), which is based on Boltzmann general expression in equation (1).

$$S\,\mathrm{Pr}ob=\exp\left(\frac{-SP\ x\ SDT}{LDT}\right). \quad (3)$$

Where;

(SDT): Shortest Delay Time (Best Time), which is a dynamic parameter that varies with number of iterations (IRN).

(LDT): Longest Delay Time (Worst Time), which is a dynamic parameter that corresponds to Temperature in Boltzmann Selection (BS) and varies with Population exploration. Equation (3) can be rearranged as in equation (4)

$$S\,\mathrm{Pr}ob=\exp\left(-SP\left(\frac{SDT}{LDT}\right)\right). \quad (4)$$

SDT and LDT are important parameters in the genetic algorithm computations, as they are related to the used fitness function. They are vital for the designed algorithm in determining the value of SProb, and thus, affect its controlling mechanism and number of Passing Vehicles (PV). Hence, SProb is a function of both the Selection Pressure (SP) and the dynamic ratio of the genetic algorithm times that depends on the population value, which is related to other parameters such as offsprings, mutants and crossover levels.

Thus, selection of SP and Pop, will affect SProb, which in turn affects the algorithm computations and Vehicles Passing Rate (VPR). So, balancing and optimization of the two values will result in a tuned, balanced and efficient GA for traffic signals and traffic control and management.

Two conditions associated with Selection Pressure:

1) High Selection Pressure Value: Early convergence.
2) Low Selection Pressure Value: Late convergence.

Thus, an optimum selection pressure value is needed for optimal control of traffic lights, which is a dynamic, changeable value function of traffic density, which is computed in the designed algorithm as in equations (5) and (6) for each road i.

$$TP_i=\left(RC_i-VP_i\right). \quad (5)$$

$$TPR_i=\left(\frac{\left(RC_i-PV_i\right)}{RC_i}\right). \quad (6)$$

Where;

TP: Traffic Pressure

TPR: Traffic Pressure Ratio

PV: Passing Vehicles

RC: Road Capacity

i: Road at the 4-road intersection (i=1, 2, 3, 4)

PV is related to both SP and Pop, subsequently is affected by SProb, which is a function of SDT and LDT and SP. Equal PV for same TP per road should be equal for the GA algorithm to be considered balanced and optimized. The general process of optimization is shown in Fig. 2.

The idea is to optimize the GA performance using both SP and Pop to match the balancing data shown in Table I.

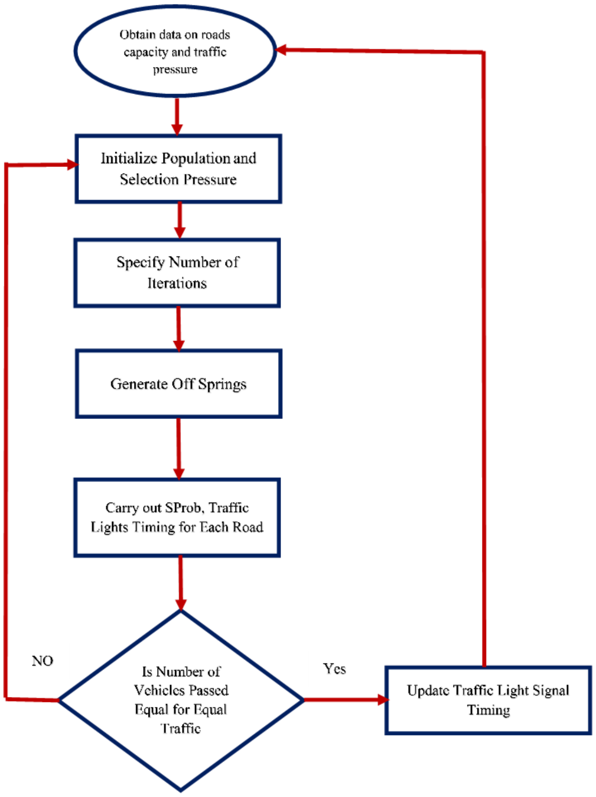

Obtain data on roads capacity and traffic pressure
Initialize Population and Selection Pressure
Specify Number of Iterations
Generate Off Springs
Carry out SProb, Traffic Lights Timing for Each Road
Is Number of Vehicles Passed Equal for Equal Traffic
NO
Yes
Update Traffic Light Signal Timing

Fig. 2. GA Optimization Algorithm.

TABLE. I. GA BALANCING REFERENCE VALUES

POP$_i$	Vehicles Passing Times (sec)				No. of Passing Vehicles			
	Rd.1	Rd.2	Rd.3	Rd.4	Rd.1	Rd.2	Rd.3	Rd.4
SP								
2	17	57	57	17	7	23	23	7
4	17	57	57	17	7	23	23	7
6	17	57	57	17	7	23	23	7
8	17	57	57	17	7	23	23	7
10	17	57	57	17	7	23	23	7

IV. RESULTS

Table II shows the number of Iterations (IRN) required for each Selection Pressure Value to achieve balanced and optimized GA performance as a function of the relationship between Population (Pop) and Selection Pressure (SP), with plots representing the data shown in Fig. 3 as a function of SP and in Fig. 4 as a function of Pop.

TABLE. II. IRN AS A FUNCTION OF BOTH SP AND POP

Iterations (IRN)	Population (Pop)				
SP	200	400	600	800	1000
2	650	350	300	50	50
4	500	200	50	150	50
6	300	50	200	50	50
8	450	50	200	50	50
10	950	400	400	400	50

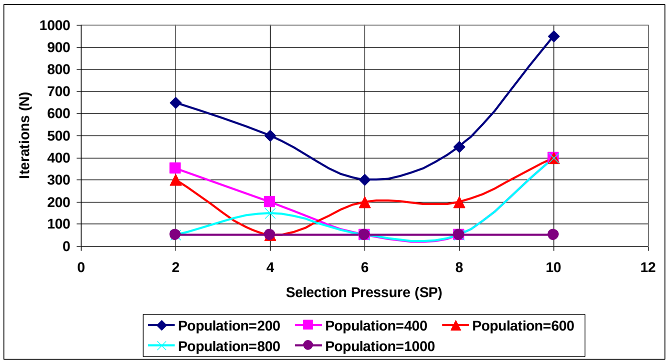


Fig. 3. Iterations (IRN) as a Function of SP Per Specific Pop.

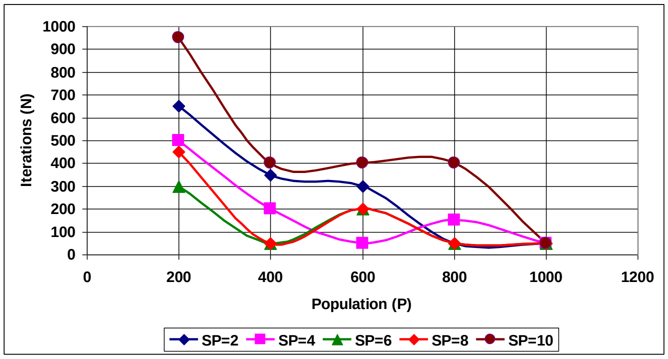


Fig. 4. Iterations (IRN) as a Function of Pop Per Specific SP.

V. ANALYSIS AND DISCUSSION

The analysis process is based on the stability criteria, whereby, a correlation between three parameters (SP, Pop, Iterations) is carried out, with final correlation to smooth and correct traffic lights control at the junction, such that the four roads will pass the correct number of vehicles. This is a calibration and load balancing measure, which is based on initial conditions of equal traffic pressure for opposing roads.

Table II shows data represent the Iteration parameter as a function of the relationship between SP and Pop, whereby Iterations are dependent on the two parameters (SP, Pop) with each SP level is fixed for a variable levels of Pop. The general form of dependency is described by equation (7).

$$IRN = f(SP, Pop). \tag{7}$$

Now, Pop affects LDT, hence affecting SProb. Also, SP affects SProb according to equation (4). Thus, IRN, will affect SProb and by using both equations (4) and (7), equation (8) is obtained.

$$S\,\mathrm{Pr}ob = f\,(IRN). \tag{8}$$

The PV parameter depends on SProb, hence, can be described by equation (9).

$$PV = g\,(S\,\mathrm{Pr}ob) = g(f(IRN)). \tag{9}$$

Then, for each opposite and equal TP roads, the difference in the rate of change should be zero for balanced GA algorithm. Thus; for i (1 to 4):

$$\Delta\, PV_i = \Delta\, g_i\left(f\left(IRN_i\right)\right) = 0.. \tag{10}$$

The condition in equation (10) will only apply if the number of iterations in the GA algorithm is the same for each SP and Pop to enable same number of vehicles to pass on each road as a function of TP. This condition is valid for stable GA algorithm. Hence, when achieved, the GA algorithm is balanced and optimized as equation (11) shows.

$$\Delta\, IRN_i = 0\,. \tag{11}$$

From Table II, it is clear that the Steady State (SS) values is reached for the GA algorithm controlling four traffic lights with one intersection at Pop=1000, with minimum number of Iterations of 50, which is also constant for all levels of SP. Such convergence, appears in Fig. 2.

Table III, presents data that shows effect of Pop on IRN in order to reach SS, with Fig. 3 showing the convergence to an SS described by the pairing in equation (12):

$$SS = (Pop, IRN)\,. \tag{12}$$

From Table II, it is realized that for some SP, Pop values oscillation occurs during tuning and balancing of the GA algorithm. The two non-oscillatory values are shown in Table III, with SP=2 having early convergence to minimum number of iterations. Non-oscilatory SP, Pop pairs are prefered in the balancing process to other values, eventhough the convergence condition achived with intermediate GA oscillations.

Fig. 5 to 8 show SDT and LDT for the two non-oscillatory SP values.

TABLE. III. NON-OSCILLATORY SP AND POP

Iterations (IRN)	**Population (Pop)**				
SP	**200**	**400**	**600**	**800**	**1000**
2	**650**	**350**	**300**	**50**	**50**
10	**950**	**400**	**400**	**400**	**50**

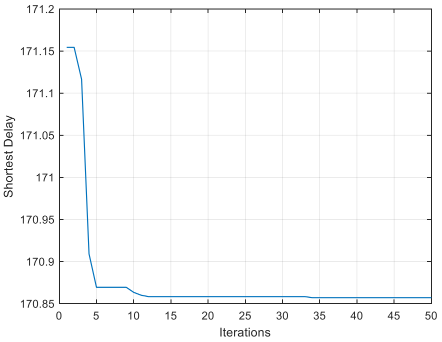


Fig. 5. {2,1000} SDT Curve.

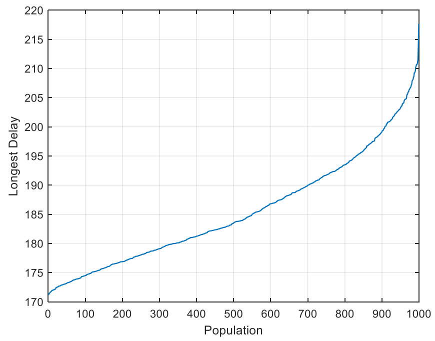


Fig. 6. {2,1000} LDT Curve.

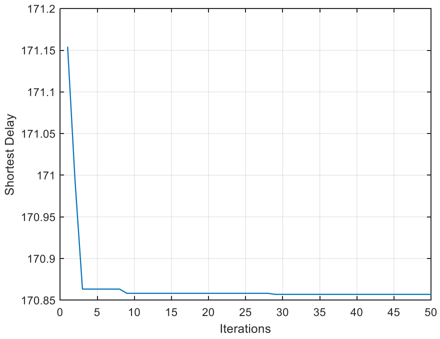


Fig. 7. {10,1000} LDT Curve.

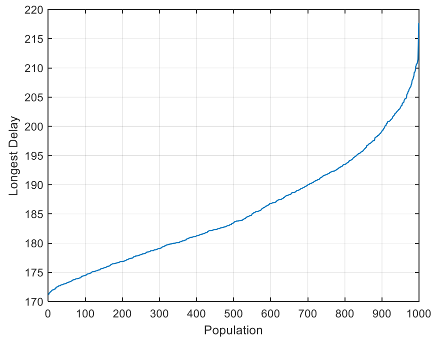


Fig. 8. {10,1000} LDT Curve.

VI. CONCLUSIONS

The obtained condition for number of interactions (IRN) in this work and subsequent steady state condition specified, is an important achievement, as it uncovers conditions of constant and low number of iterations necessary to obtain balanced, and optimized GA control algorithm used in the management of traffic lights signaling process. The used technique locally weights the processes using both Pop and SP correlation, which enables optimized functionality and better forecasts to achieve optimal dynamic traffic modeling. The balanced GA algorithm proposed in this work offers tangible advantages. Future work requires the extension of such application of balanced GA algorithm is recommended to cover multi-section traffic control, whereby each four-road, single intersection is regarded as a single parameter and correlated with other similar structure to cover more complex arrangement.

REFERENCES

[1] J. Chen, Y. Yu, Q. Guo , “Freeway Traffic Congestion Reduction and Environment Regulation via Model Predictive Control,” Algorithms, vol. 12, no. 220, pp. 1–23, 2019.

[2] G. Jia , R. Ma, Z. Hu , “Review of Urban Transportation Network Design Problems Based on CiteSpace,” Journal of Theoretical and Applied Information Technology, vol. 2019, ID. 5735702, pp. 1–22, 2019.

[3] A. Goyal, M. Singh, A. Aeron, “Simulation of Traffic Optimization to Reduce Congestion,” International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 11, pp. 3780–3783, 2019.

[4] S. Yang, Y. Ji, D. Zhang, J. Fu, “Equilibrium between Road Traffic Congestion and Low-Carbon Economy: A Case Study from Beijing, China,” sustainability, vol. 2019, no. 11, pp. 1–22, 2019.

[5] B. Sony, A. Chakravarti, M. Reddy, “Traffic Congestion Detection using Whale Optimization Algorithm and Multi Support Vector Machine,” Journal of Theoretical and Applied Information Technology, vol. 7, no. 6C2, pp. 589–593, 2019.

[6] Q. Bing , D. Qu , X. Chen , F. Pan, J. Wei, “Short-Term Traffic Flow Forecasting Method Based on LSSVM Model Optimized by GA-PSO Hybrid Algorithm,” Discrete Dynamics in Nature and Society, vol. 2018, ID. 3093596, pp. 1–10, 2018.

[7] Y. Chen L. Rilett, “Signal Timing Optimization for Corridors with Multiple Highway-Rail Grade Crossings Using Genetic Algorithm,” Journal of Advanced Transportation, vol. 2018, ID. 9610430, pp. 1–14, 2018.

[8] N. Saharkar , M. Wanjari, “A Genetic Algorithm Based Approach to Solve Transport Problems for School Buses?” Journal of Engineering and Applied Sciences, vol. 13, no. 4, pp. 848-851, 2018.

[9] T. Karthy and K .Ganesan, “Multi Objective Transportation Problem - Genetic Algorithm Approach,” International Journal of Pure and Applied Mathematics, vol. 119, no. 9, pp. 343-350, 2018.

[10] E. Han, H. PiLee, S. Park, J. So, I. Yun, “Optimal Signal Control Algorithm for Signalized Intersections under a V2I Communication Environment,” Atmoshphere, vol. 2019, ID. 6039741, pp. 1–9, 2019.

[11] A. Potnurwar, S. Aote, V. Bongirwar, “Design of Traffic Volume Forecasting based on Genetic Algorithm,” International Journal of Recent Technology and Engineering, vol. 8, no. 2, pp. 4264–4268, 2019.

[12] A. EL Idrissi, C. Tajani, M. Sabbane, “New Crossover Operator for Genetic Algorithm to Resolve The Fixed Charge Transportation Problem,” Journal of Theoretical and Applied Information Technology, vol. 98, no. 8, pp. 1607–1617, 2017.

[13] W. Wen-jing, “Improved Adaptive Genetic Algorithm for Course Scheduling in Colleges and Universities,” iJET, vol. 13, pp. 29–42, 2018.

[14] Y. Wang , X. Yang , H. Liang , Y. Liu, “A Review of the Self-Adaptive Traffic Signal Control System Based on Future Traffic Environment,” Journal of Advanced Transportation, vol. 2018, ID. 1096123, pp. 1–12, 2018.

[15] Q. Tang, B. Friedrich, “Design of Signal Timing Plan for Urban Signalized Networks including Left Turn Prohibition,” Journal of Advanced Transportation, vol. 2018, ID.1645475, pp. 1–16, 2018.

[16] S. Srivastava, S. Sahana, “Application of Bat Algorithm for Transport Network Design Problem,” Applied Computational Intelligence and Soft Computing, vol. 2019, ID. 9864090, pp. 1–12, 2019.

[17] Z. Cakici, Y. Sazi Murat, “A Differential Evolution Algorithm-Based Traffic Control Model for Signalized IntersectionsAdvances in Civil Engineering, vol. 2019, ID. 7360939, pp. 1–16, 2019.

[18] C. Canali, R. Lancellotti, “GASP: Genetic Algorithms for Service Placement in Fog Computing Systems,” Algorithms, vol. 12, pp. 1–19, 2019.

[19] A. Rahman, N. Shahruddin, Ismail Ishak, “Solving the Goods Transportation Problem Using Genetic Algorithm with Nearest-Node Pairing Crossover Operator, ” Journal of Physics: Conference Series, vol. 1366, pp. 1–9, 2019.

[20] X. Feng, X. Zhu, X. Qian, Y. Jie, F. Ma, and X. Niu, “A new transit network design study in consideration of transfer time composition,” Transportation Research Part D: Transport and Environment, vol. 66, pp. 85–94, 2019.

A New Approach for Multi-Level Evaluation of Strategic Educational Goals

Mohammad Alhaj[1], Mohammad Hassan[2], Abdullah Al-Refai[3]
Computer Engineering Dept., Faculty of Engineering
Al-Ahliyya Amman University, Amman, Jordan

***Abstract*—Educational organizations with multiple level of management promotes for their strategic educational goals as a correlated and clustered data. The typical assessment and feedback approaches are paper-based where word documents and flowcharts are used to evaluate strategic educational goals augmented with quantitative indicators. Unfortunately, the paper-based approach often neglects the relationship and dependencies between the educational goals defined at different levels. This may lead to complications in the analysis, lack of clarity, and subject to different interpretations by the multiple management. We propose a multi-level model-driven approach that improves the assessment of strategic educational goals, handles the clustered data efficiently and allows the individual and group level assessment to take effect simultaneously. The approach also allows decision makers in academic institution to extract valuable information from goal models at different academic levels and measure the fulfilment of the educational goals with respect to the target performance in a formal way.**

Keywords—Evaluation process; goal model; multi-level modelling; goal requirement language; program educational goals

I. Introduction

Academic institutions are using different learning assessment approaches and reviews to evaluate student’s learning progress. The evaluation process of students’ performance starts by the time they are admitted to the academic institution and continues until four or five years from the graduation when they are engaged with the market.

Strategic educational goals are those objectives and targets that support performance roadmap to measure the institution state and progress. Academic institutions promote for strategic educational goals at six academic levels as in Fig. 1. The top level is the institution where the vision statement is used to describe the future accomplishments and objectives of the institution; also, the mission statement to describe the action needed to be done to meet the vision statement. At the faculty and department levels, mission statements stem their targets from the institution mission. At the program level, the program educational objectives (PEOs) are tailored to serve and promote the mission statement to describe the professional and career accomplishments of graduates during the four to five years from graduation [1]. Finally, the student outcomes (SOs) are defined at the curriculum and course levels, to describe what students are expected to know and practice by the time of graduation from the program.

Several constituencies are involved in the development and evaluation process of strategic educational goals. The major constituencies are program academic members, industrial advisory board (IAB), program alumni, undergraduate students of the program, and employers of the program graduates. Other constituencies may be involved are students’ parents, program administrative staff and administration of the academic institution. IAB consists of professionals, experts and/or managers employed at major industries related to the academic program.

Student’s performance during the academic semester is used as an indicator of how much the SOs have been met. There are two types of assessment tools are used in measuring student performance: a) direct tools where student assessments are measured though direct examination or various of submitted work, such as assignments, quizzes and exams; b) indirect tools where student achievement requires that academic consistencies infer actual student abilities, knowledge, and values rather than observe direct evidence of achievement, example of indirect tools: surveys and interviews.

The typical paper-based learning assessments and reviews approaches, nowadays, comes in a form of word documents, spreadsheets and flowcharts. They are used in the process of evaluating strategic educational goals with respect to the target goals and objectives of the institution. The paper-based approach often neglects the relationship and dependencies between the educational goals defined at different levels. This might cause confusion in analyzing the learning assessments, lack of clarity, and subject to different interpretations by constituencies. It is desirable to use a model-driven approach with multi-level modeling to improve the learning assessment, evaluate the learning goals in a formal way and extract information at different academic levels.

The proposed paper extends additional details on earlier research results presented at the conference in [1]. In this paper, we proposes a model-driven approach with a multi-level modelling where the Goal-oriented Requirement Language (GRL) is used for assessing the learning goals and objectives. Multi-level Goal modeling provides performance indicators for the quantitative measures of strategic educational goals during the continuous evaluation process. It handles the collective data efficiently and allows the individual and group level assessment to take effect simultaneously. The paper is organized as follows: Section 2 presents the background and related work; Section 3 describes an overview of our goal-oriented approach; Section 4 demonstrates the GRL goal modeling; Section 5 shows a case study of evaluation modeling and analysis; Section 6 provides conclusions and future work.

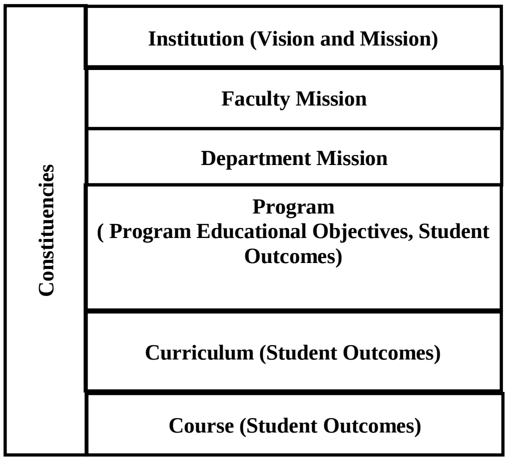

Constituencies
Institution (Vision and Mission)
Faculty Mission
Department Mission
Program
(Program Educational Objectives, Student Outcomes)
Curriculum (Student Outcomes)
Course (Student Outcomes)

Fig. 1. Strategic Educational Goals of an Academic Institution.

II. Background and Related Work

Goal-oriented modeling languages are used in recent research projects to capture business goals and associate them with performance measures on different quality aspects. The aim of modeling is to improve the decision-making process, provide a structure formality, reduce the lack of clarity in user requirements and detect early of any deficiencies in meeting business goals. The Goal-oriented Requirement Language (GRL) is a standard notation for goal modeling. It is part of the User Requirements Notation (URN) [2] that describes business goals and facilitates the modelers to describe intention elements (e.g., goals, tasks, indicators) their decomposed structure (e.g., sub goals, stubs), connecters (e.g., dependencies, contribution) and their corresponding partners (e.g., actors, agents, teams).

In many recent researches, multi-level goal modeling is used when business goals are organized at more than one level. Multi-level goal modeling is hierarchical structured business goal that allows researchers to investigate the effect of group attributes on individual business goal while accounting for non-independence of observations. The analysis of multi-level goal modeling at the lower level is performed to individual business goals and nested within the accumulated business goals at the higher level [3]. Multi-level modeling is used when the analyzed data have a clustered structure and there is a substantive interest in the individual effects, group effects and the mutual effects. Multi-level modeling was adapted in different researches, such as public health [4] and validating education indicators [5].

As part of the proposed approach, a common open source graphical editor called jUCMNav [6] is used for goal modeling. jUCMNav is an eclipse-based URN tool that supports modeling goals and business processes with GRL and Use Case Map (UCM). It enables generating and managing complex GRL models based on multi-level modeling. It also provides features to utilize strategies using different analysis algorithms, to support execute and visualize analysis results, and to generate reports.

Research projects are using multi-level approaches for different evaluations. Hoe et al. in [7] propose an evaluation framework for usability of a mobile phone using a multi-level hierarchical model of usability factors. Sanders in [8] reviews the development progress of performance/dependability evaluation tools, and the importance of creating modeling frameworks that support multi-level modeling and multiple solution methods as an integrated framework. While Comuzzi et al. in [9] present the fundamental elements and interfaces of the technical architecture for a multi-level SLA management framework. Also, Yang and Sen in [10] develop a general multilevel evaluation process that deals with multiple attribute decision making problem with both quantitative and qualitative attributes.

Several research projects also have been using GRL modeling language for business goal compliance. Tawhid et al. in [11] propose a novel approach that models regulations with the GRL enhanced with qualitative indicators to generate questions for inspection operations and facilitate compliance analysis. A framework in [12] uses metrics defined in goal and scenario models to validate quality assurance of online business processes. Ghanavati et al. in [13] propose a framework that models legal documents with goals and maps such model to the goal model of the organization. To analyze the degree of legal compliance/non-compliance of organizational goals, traceability links are used between these two models and the GRL quantitative and qualitative algorithms.

Also, several learning assessment and feedback approaches and methodologies have been followed recently. The target is evaluating the learning outcomes and objectives of academic institution. Suskie in [14] introduces and analyzes various methods and approaches of assessing student learning outcomes. Gastli et al. [15] propose an innovative tool and process that allows accurate direct and indirect outcomes assessment of courses and programs while facilitating the tasks for the instructors and in their evaluation process. DeLyser and Hamstad in [16] discuss the visit made by the ABET team to review the outcomes assessment process at University of Denver and what changes were and are continually being made. Yue in [17] proposes a course-based approach that associates learning outcome objectives with accreditation standards and courses; a suitable assessment tool can then be used to assess the course. Besterfield-Sacre et al. [18] develop a framework that specifies the learning outcomes of engineering faculty by expanding them into a set of attributes.

In summary, it is obvious that the works above have addressed some features that are similar to our work. The major advantage of the proposed approach compared to the others is that it integrates the features of using model-driven engineering with the multi-level architecture to build an approach for evaluating the objectives and goals of an academic institution. The proposed approach is used in to improve the assessment of strategic educational goals, handles the clustered data efficiently and allows decision makers in academic institution to extract valuable information from goal models at different academic levels and measure the fulfilment of the educational goals with respect to the target performance in a formal way.

III. An Overview of the Proposed Approach

In this paper, we propose a multi-level evaluation approach that supports a substantive interest in the individual effects, group effects and the mutual effects of educational. The approach helps avoiding any unintentional complexities with multi-level goal modelling. This allows decision makers in academic institution to extract information from goal models at different academic levels, discovers patterns in large volume of details and investigate the effect of group attributes on individual business goal while accounting for non-independence of observations. The evaluation approach may occur at six nested academic levels depending on the structure of the academic institution and granular details that are desired, as in Fig. 1. It starts at the course and/or curriculum at the bottom-level, the program and department at the middle-level and ends with the faculty and institution at the top-level when accumulated reviews are desired.

The general view of the multi-level evaluation approach of strategic educational goals is described in Fig. 2. It consists of five steps used at each level:

1) The Definition step is used to identify preliminary details, such as the strategic educational goals under evaluation, the target constituencies responsible for evaluation and collecting data, the direct/indirect assessment tools that will be used and the collection frequency or period of the evaluation.

2) The Assessment step provides two types of assessment techniques to measure the performance metrics of the strategic education goals: a) the paper-based, where documents and spreadsheets are used as a scoring guide, e.g., Rubric [19] and CAP [20]; and b) model-based where software modelling is used for goal assessment, e.g., GRL and i* [2].

3) The Evaluation step, where the output of the assessment techniques is accumulated to the next level of assessment and also is used to develop a set of recommendations for improvement.

4) The Adoption step, where the constituencies deal with two types of recommendations, the short-term recommendation which may be refined in order to meet the regulations and bylaws and adopted during the semester. Also, there is the long-term recommendation which may take a further discussion and approvals.

5) The Implementation step, where the adopted recommendations are implemented, and constituencies are informed of significant improvements during the public forums. The processes return to the assessment and repeat for another cycle.

Fig. 3 describe the evaluation process of the proposed approach. The course evaluation process is a bottom-level where assessment seeks input from individual faculty members for each course taught. A course has specific student learning outcomes (SOs) designed to achieve a number of selected learning attributes such as knowledge solving problems, communication skills and leadership. Generally, the direct assessment results of the exams, quizzes, assignments, etc. are reflected to the grades. Based on the performance results of PEOs and SOs, a detailed summary of the improvement introduced in the course report; and a set of recommendations to be approved by a focus group consisting of faculty members who are considered in the knowledge field of that course.

At the curriculum evaluation process, teams of focus groups accumulate the individual contribution from all courses to the curriculum PEOs and SOs in order to assess the contribution of the entire curricula. This helps to minimize inconsistencies in teaching the courses by different faculty members and have a coherent structure of courses in the area.

At the program evaluation processes, the department council, all instructors and part of students are involved. Inputs are obtained through direct interaction with individual students and academic advisors are discussed in the department council and changes, corrective/preventive actions are proposed. The Proposed changes are discussed at the program level. Changes that do not conflict with the institution regulations and bylaws are implemented. While other actions that may conflict are submitted in a form of a proposal to the department council who may take action or may forward them to the Faculty Deans' Council for further discussion.

At the top-level management evaluation process, different councils accumulate the results of every program in order to assess the contribution of the entire programs, and the performance results of the PEOs that are reflected on the department mission.

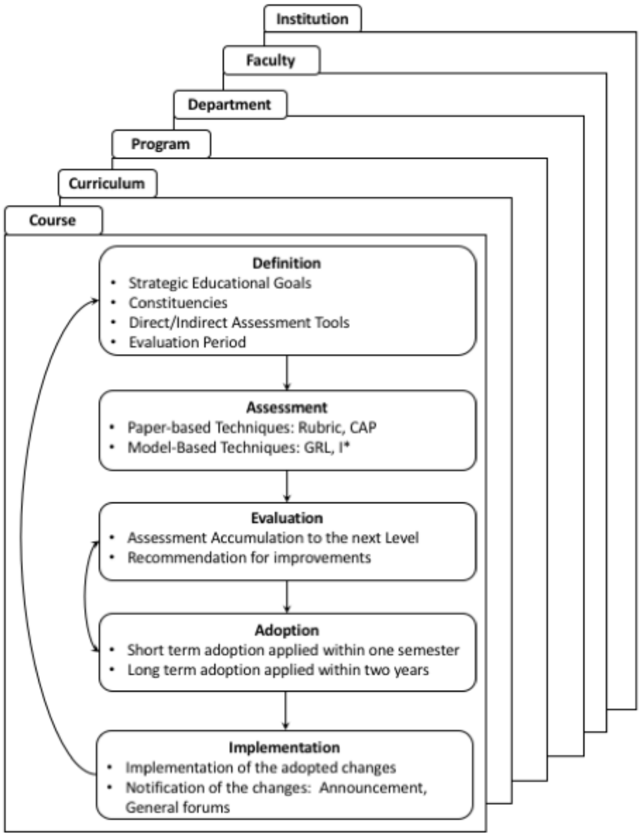

Institution
Faculty
Department
Program
Curriculum
Course
Definition
• Strategic Educational Goals
• Constituencies
• Direct/Indirect Assessment Tools
• Evaluation Period
Assessment
• Paper-based Techniques: Rubric, CAP
• Model-Based Techniques: GRL, I*
Evaluation
• Assessment Accumulation to the next Level
• Recommendation for improvements
Adoption
• Short term adoption applied within one semester
• Long term adoption applied within two years
Implementation
• Implementation of the adopted changes
• Notification of the changes: Announcement, General forums

Fig. 2. Multi-Level Evaluation of Strategic Educational Goals.

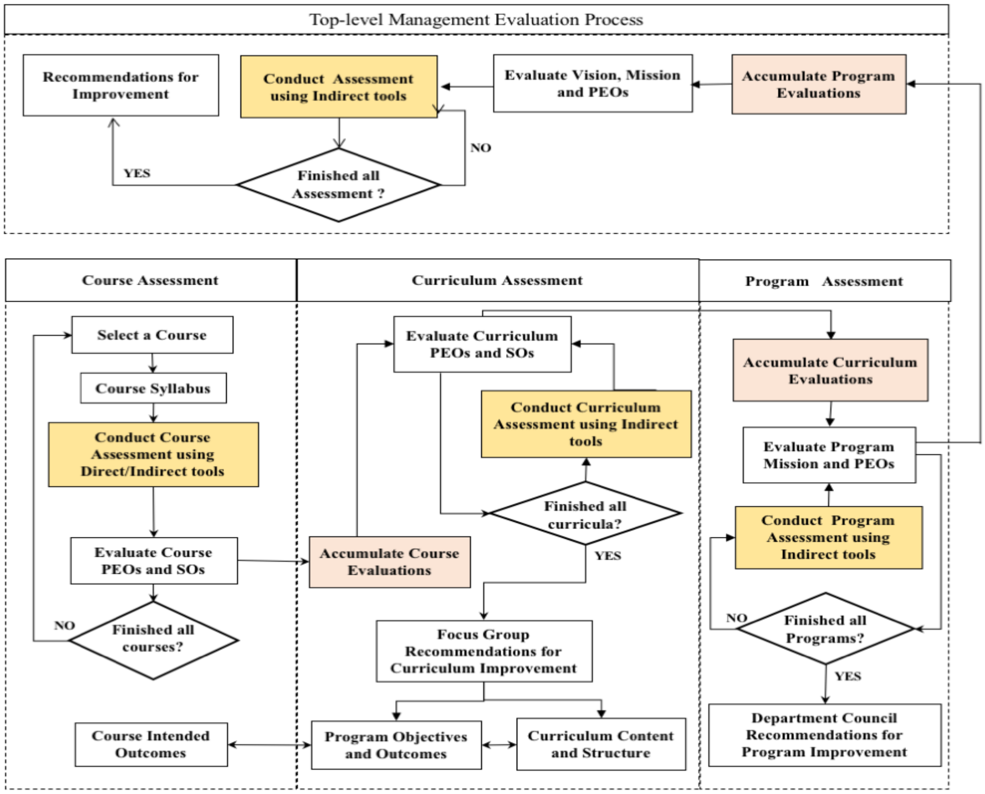

Top-level Management Evaluation Process
Recommendations for Improvement
Conduct Assessment using Indirect tools
Evaluate Vision, Mission and PEOs
Accumulate Program Evaluations
NO
YES
Finished all Assessment ?
Course Assessment
Curriculum Assessment
Program Assessment
Select a Course
Course Syllabus
Conduct Course Assessment using Direct/Indirect tools
Evaluate Course PEOs and SOs
NO
Finished all courses?
Course Intended Outcomes
Evaluate Curriculum PEOs and SOs
Conduct Curriculum Assessment using Indirect tools
Finished all curricula?
YES
Accumulate Course Evaluations
Focus Group Recommendations for Curriculum Improvement
Program Objectives and Outcomes
Curriculum Content and Structure
Accumulate Curriculum Evaluations
Evaluate Program Mission and PEOs
Conduct Program Assessment using Indirect tools
NO
Finished all Programs?
YES
Department Council Recommendations for Program Improvement

Fig. 3. The Evaluation Process at Six Levels of Assessments.

Another step at the top-level management process is the faculty and institute evaluation process, where the accumulated performance results of the department missions are used as an assessment measures of the faculty mission. Finally, the accumulated performance results of the faculties' missions are used for assessment measures of the institute mission and vision. A set of recommendations is made by the faculty council or institute council for further improvements that comply with the institute mission and vision.

IV. Using Multi-level Goal Modelling Technique

In the proposed approach, a GRL modelling language was used to assess the strategic educational goals using goal models. The GRL goal model is described as part of URN language and is supported by jUCMNav [6]. Each element in the proposed data model is mapped into model element in the multi-level goal models. The high-level goals Vision and Mission are mapped into GRL Softgoal ; while the PEOs and SOs are mapped into the sub-goals . Model elements are linked together respectively and each one of them may contribute fully or partially at different assessment levels. The contribution relationship describes how an element participate to the other elements in GRL model; the contribution value ranges from -100% (negative), 0% (neutral) to +100% (positive).

At the course and curriculum assessment levels, courses that are selected for assessment are mapped into resources . Assessment tools are also mapped into the goal model as key performance indicators (KPIs) . KPI contains the constituency's achievement provided by the direct/indirect assessment tools and contributes to different strategic educational goals in the model. A KPI has an evaluation value that measures the current situation. It ranges between the -100% (negative), +0% (neutral) or +100% (positives) values. Constituencies are represented as actors to define their ownership and responsibility in the goal model.

Fig. 4 defines an arbitrary multi-level goal model at two assessment levels. At the top-level goal model, two program educational objective PEO1 and PEO2 contribute to a Mission by 50% and 30% respectively; and at the higher level the Mission contributes to the institute Vision by 100%. There are also two indirect assessment tools (KPIs): Indirect Assessment2 and Indirect Assessment3 contribute both to the Mission by 10% and 20%, respectively. At the bottom-level goal model, a sample Course of the curriculum, represented as a resource contributes to two program educational objectives PEO1 and PEO2 by 25% and 50%, respectively.

To maintain the correlation between the multi-level goal model, traceability links are used between the PEO1 and PEO2 at the top-level and bottom-level GRL models; such that changes in their evaluation values at any GRL model will be reflected to the other models. Two student outcomes SO1 and SO2 contribute both to the Course by 75% and 60% respectively. There are also three assessment tools (KPIs): Direct Assessment1, Direct Assessment2 and Indirect Assessment1 are used for evaluating the student outcomes SO1 and SO2. They contribute by 10%, 25%, 15% and 20% respectively.

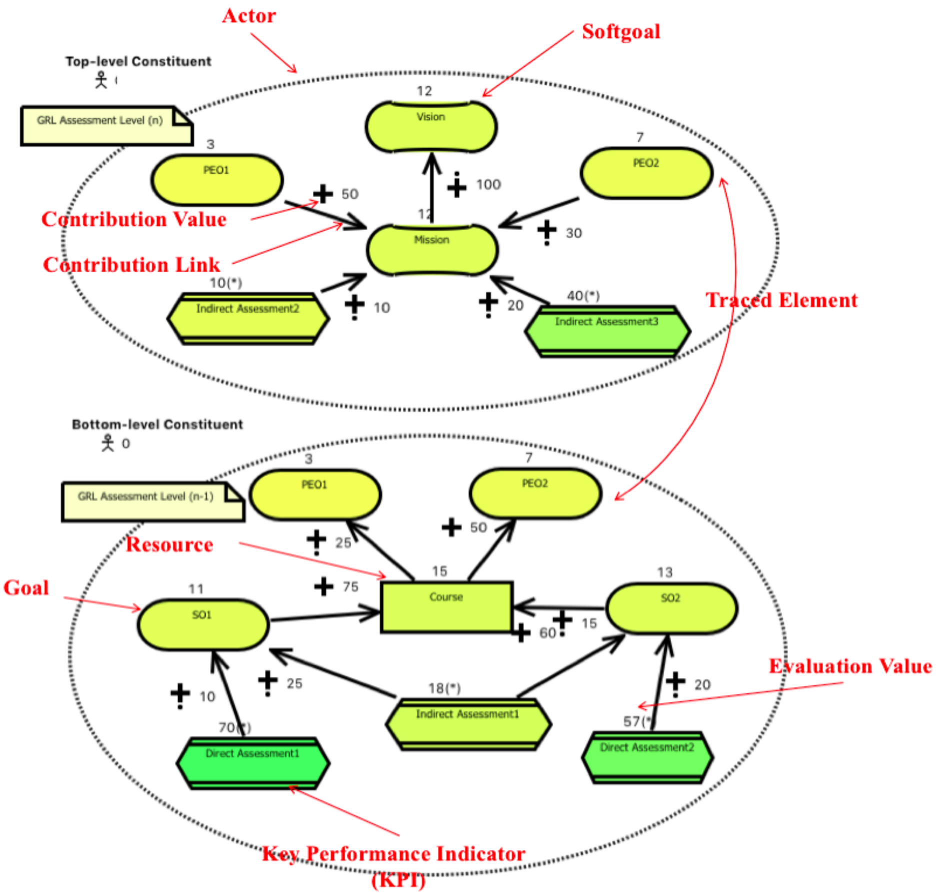


Fig. 4. An Arbitrary Multi-Level GRL Model.

V. A Case Study of the Proposed Approach

We introduce a case study developed at Al-Ahliyya Amman University (AAU) [21], where 10 faculties were involved with a total of 38 departments and 50 programs. Three curriculum used for evaluations: fall, winter and summer with an average of 55 courses used for assessment for each program.

At the Definition step, the participation of the strategic educational goals, constituencies, assessment tools at different levels and an approximate evaluation period is described in Table I. The elements of strategic educational goals are involved such that the vision and mission reflect the objectives at the high levels and PEOs and SOs are used to measure the performance metrics at the low levels. Ideally, all constituencies need to participate actively, however, this ideal approach may not be achieved with ease and become consistent all the times.

The indirect assessment tools are used to evaluate the satisfaction of the constituencies that is reflected on the performance results. Different kind of surveys are used for that purpose with a motive of assessing the institution’s regulations, polices and activities with respect to its vision, mission, PEOs and SOs. The common surveys used in academy are trend surveys, panel surveys, cohort surveys [22]. The direct assessment tools are used only at the course level using tools such as assignments, quizzes and exams. Periodic assessment of strategic educational goals varies between levels. It is defined by the participated consistencies during the evaluation process as in Table I.

At the Assessment step, a sample of the case study models developed at the Civil Engineering Department in the Faculty of Engineering as in Fig. 5-10. The goal model, in Fig. 5, describes the strategic educational goals at institution level. The Deans Council is the major actor that contains a sample of four Faculties’ Missions, each one contributes to the AAU Mission by 25%; while AAU Mission contributes to AAU Vision by 100%. There are also three indirect assessment tools (surveys), each contributes to AAU Mission by 20%. These tools measure the satisfactions of random constituencies with respect to the provided services.

The goal model below, in Fig. 6, describes the strategic educational goals at faculty level. The Faculty Council is the major actor that contains the mission of five engineering departments: computer engineering, civil engineering, communications and electronics engineering, electrical engineering and medical engineering, each one contributes to the Faculty of Engineering Mission by 20%. A traceability link between the Faculty of Engineering Mission in Fig. 6 and its equivalent in upper goal model as in Fig. 5. Two indirect assessment tools (surveys) are also used to measure the satisfaction of provided services at the faculty level. Each survey contributes to the Faculty Mission by 10%.

TABLE. I. THE DEFINITION STEP OF THE CASE STUDY

Evaluation Level	**Strategic Educational Goals**	**Constituency**	**Assessment Tool**		**Approximate Evaluation Period**
			Direct	**Indirect**	
Institute	Institution Vision and Mission, Faculty Mission	Institution Council, Deans Council	NA	√	From 5 to 7 Years
Faculty	Faculty Mission, Dept. Mission	Faculty Council	NA	√	From 3 to 5 years
Department	Dept. Mission and PEOs	Department Council, Industrial Advisory Board, Alumni,	NA	√	Every 2 years
Program	PEOs	Department Council, Industrial Advisory Board, Alumni, Parents	NA	√	Every 2 years
Curriculum	PEOs and SOs	Focus Group, Enrolled Students, Graduated Students	NA	√	Every semester
Course	PEOs, SOs	Instructor, Enrolled Students	√	√	Every semester

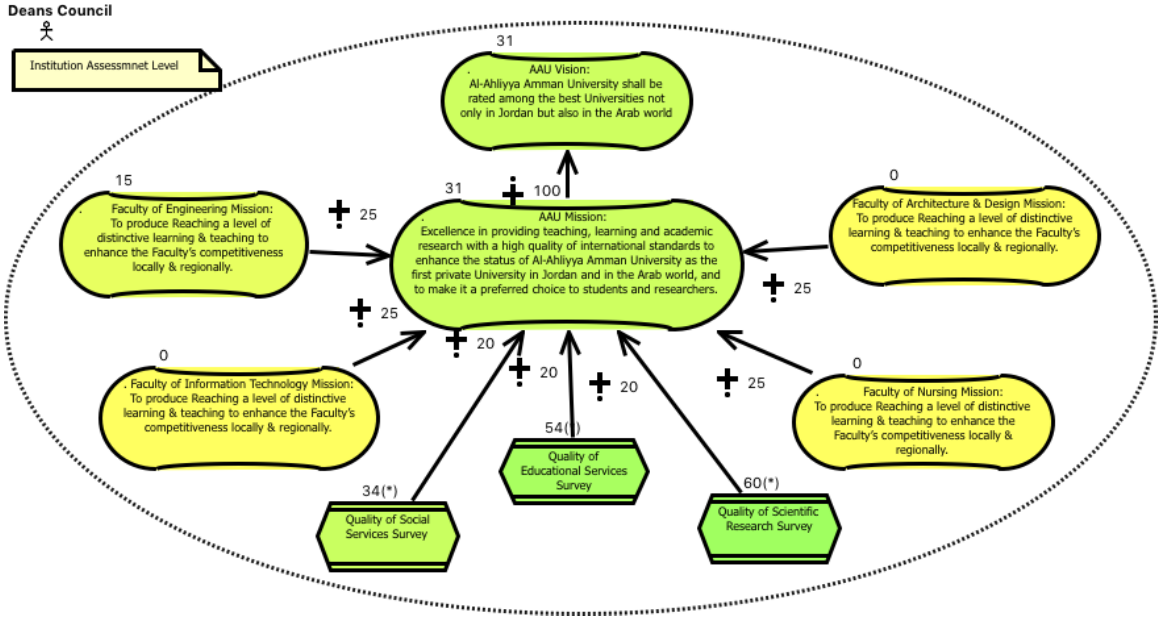

Deans Council
Institution Assessmnet Level
31
AAU Vision:
Al-Ahliyya Amman University shall be rated among the best Universities not only in Jordan but also in the Arab world
100
31
AAU Mission:
Excellence in providing teaching, learning and academic research with a high quality of international standards to enhance the status of Al-Ahliyya Amman University as the first private University in Jordan and in the Arab world, and to make it a preferred choice to students and researchers.
15
Faculty of Engineering Mission:
To produce Reaching a level of distinctive learning & teaching to enhance the Faculty's competitiveness locally & regionally.
25
0
Faculty of Architecture & Design Mission:
To produce Reaching a level of distinctive learning & teaching to enhance the Faculty's competitiveness locally & regionally.
25
25
0
Faculty of Information Technology Mission:
To produce Reaching a level of distinctive learning & teaching to enhance the Faculty's competitiveness locally & regionally.
20
20
20
25
0
Faculty of Nursing Mission:
To produce Reaching a level of distinctive learning & teaching to enhance the Faculty's competitiveness locally & regionally.
54(*)
Quality of Educational Services Survey
34(*)
Quality of Social Services Survey
60(*)
Quality of Scientific Research Survey

Fig. 5. GRL Model at the Institution Level.

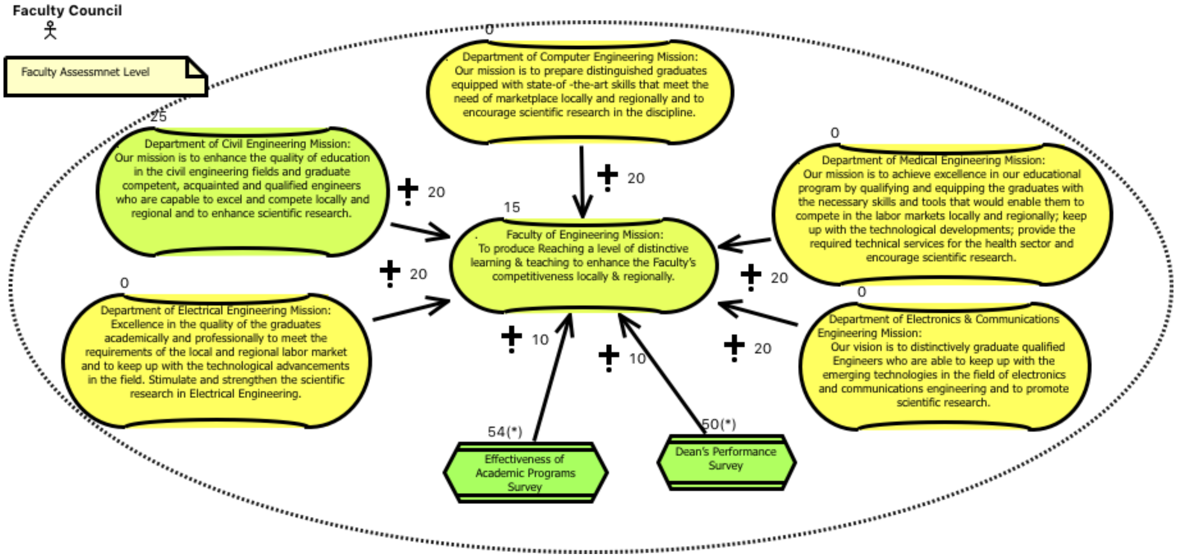

Faculty Council
Faculty Assessmnet Level
0
Department of Computer Engineering Mission:
Our mission is to prepare distinguished graduates equipped with state-of -the-art skills that meet the need of marketplace locally and regionally and to encourage scientific research in the discipline.
25
Department of Civil Engineering Mission:
Our mission is to enhance the quality of education in the civil engineering fields and graduate competent, acquainted and qualified engineers who are capable to excel and compete locally and regional and to enhance scientific research.
20
20
15
Faculty of Engineering Mission:
To produce Reaching a level of distinctive learning & teaching to enhance the Faculty's competitiveness locally & regionally.
0
Department of Medical Engineering Mission:
Our mission is to achieve excellence in our educational program by qualifying and equipping the graduates with the necessary skills and tools that would enable them to compete in the labor markets locally and regionally; keep up with the technological developments; provide the required technical services for the health sector and encourage scientific research.
20
20
0
Department of Electrical Engineering Mission:
Excellence in the quality of the graduates academically and professionally to meet the requirements of the local and regional labor market and to keep up with the technological advancements in the field. Stimulate and strengthen the scientific research in Electrical Engineering.
0
Department of Electronics & Communications Engineering Mission:
Our vision is to distinctively graduate qualified Engineers who are able to keep up with the emerging technologies in the field of electronics and communications engineering and to promote scientific research.
10
10
20
54(*)
Effectiveness of Academic Programs Survey
50(*)
Dean's Performance Survey

Fig. 6. GRL Model at the Faculty Level.

At the department level, in Fig. 7, the goal model describes the strategic educational goals of the civil engineering department. The Department Council is the major actor that contains two programs: bachelor's degree in civil engineering and master's degree in intelligent transportation systems. Each of the programs contributes to Department of Civil Engineering Mission by 70% and 30%, respectively. A traceability link is defined between the Department of Civil Engineering Mission in Fig. 7 and with its equivalent in the upper goal model as in Fig. 6. Each program defines its own PEOs. Four PEOs contributes to bachelor's degree in civil engineering program by 25%; while three PEOs contributes to master's degree in intelligent transportation systems program by 25%, 50% and 25%, respectively. Two surveys are also used to measure the satisfaction of provided services at the department level. Each survey contributes to only the master's degree program by 10%.

At the program level below, in Fig. 8, describes the strategic educational goals of the bachelor's degree in civil engineering program. The Department Council is the major actor that defines three semester curricula of the program: first, second and summer. Each curriculum contributes to bachelor's degree program by 25%. There are also four PEOs contribute to semesters' curricula defined as follows:

- PEO1: Succeed and excel in developing sound solutions to civil engineering problems.
- PEO2: Communicate and work competently in one or more core of civil engineering areas of practice or through graduate studies.
- PEO3: Work effectively and conduct themselves ethically in their professional environment and grow in their careers working on projects designed for the well-being of their society.
- PEO4: Be aware of contemporary changes and engage in life-long learning in their profession and acquire professional engineering registration.

Traceability links are defined between the bachelor's degree in civil engineering program and PEOs in Fig. 8 with their equivalents in the upper goal model in Fig. 7. Two surveys are also used to measure the satisfaction of provided services at the program level. Each survey contributes to first and second semester curriculum by 25%, respectively.

At the curriculum level below, in Fig. 9, describes the strategic educational goals of the Summer Semester Curriculum of the bachelor's degree in civil engineering program. A Focus Group in the civil engineering program is the major actor that defines two PEOs, PEO1 and PEO2, which contribute to the Summer Semester Curriculum by 30%. Traceability link is defined between the Summer Semester Curriculum in Fig. 9 with its equivalent in the upper goal model in Fig. 8. A sample of four courses are modelled: Transportation Engineering, Statics, Engineering Geology and Graduation Project (2). These courses contribute to PEO1 and PEO2 by 55%, 45% and 25%. There are also four student outcomes SOs [23] contribute to the four courses defined as follows:

- SO1: An ability to apply knowledge of mathematics, science, and engineering.
- SO2: An ability to identify, formulate, and solve engineering problems.
- SO3: Recognition of the need for, and an ability to engage in life-long learning.
- SO4: Knowledge of contemporary issues.

At the course level below, in Fig. 10, describes the strategic educational goals of the 0863300 Transportation Engineering course of the Summer Semester Curriculum. The Instructor of the course is the major actor that defines three SOs that contribute to the 0863300 Transportation Engineering course by 30%, 40% and 25%, respectively. The 0863300 Transportation Engineering course contributes to two PEOs, PEO1 and PEO2, by 45% and 55%. Traceability links are defined between the 0863300 Transportation Engineering course and the PEOs in Fig. 10 with their equivalent in the upper goal model in Fig. 9. There are also four types of direct assessment tools: Quizzes, Assignments, Midterm Exam and Final Exam, that contributes to the three SOs.

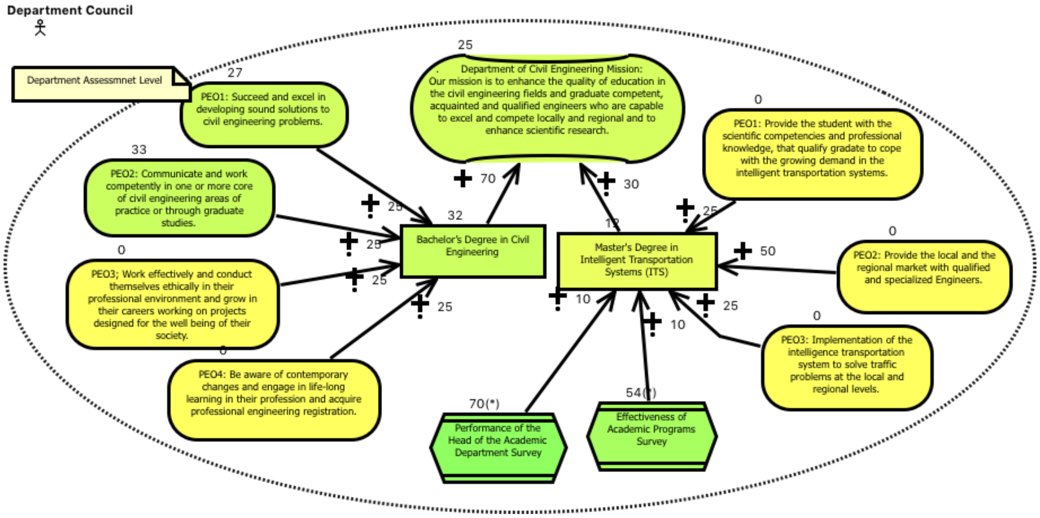

Department Council
Department Assessmnet Level
27
PEO1: Succeed and excel in developing sound solutions to civil engineering problems.
25
Department of Civil Engineering Mission: Our mission is to enhance the quality of education in the civil engineering fields and graduate competent, acquainted and qualified engineers who are capable to excel and compete locally and regional and to enhance scientific research.
0
PEO1: Provide the student with the scientific competencies and professional knowledge, that qualify gradate to cope with the growing demand in the intelligent transportation systems.
33
PEO2: Communicate and work competently in one or more core of civil engineering areas of practice or through graduate studies.
+ 70
+ 30
+ 25
+ 25
32
Bachelor's Degree in Civil Engineering
12
Master's Degree in Intelligent Transportation Systems (ITS)
+ 25
+ 50
0
PEO2: Provide the local and the regional market with qualified and specialized Engineers.
0
PEO3; Work effectively and conduct themselves ethically in their professional environment and grow in their careers working on projects designed for the well being of their society.
+ 25
+ 25
+ 10
+ 10
+ 25
0
PEO3: Implementation of the intelligence transportation system to solve traffic problems at the local and regional levels.
0
PEO4: Be aware of contemporary changes and engage in life-long learning in their profession and acquire professional engineering registration.
70(*)
Performance of the Head of the Academic Department Survey
54(*)
Effectiveness of Academic Programs Survey

Fig. 7. GRL Model at the Department Level.

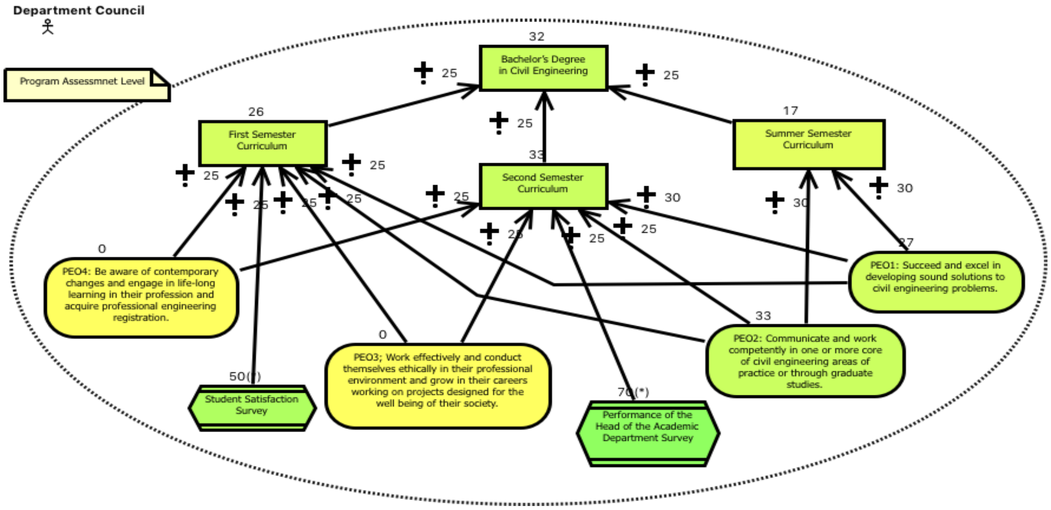

Department Council
Program Assessmnet Level
32
Bachelor's Degree in Civil Engineering
26
First Semester Curriculum
33
Second Semester Curriculum
17
Summer Semester Curriculum
0
PEO4: Be aware of contemporary changes and engage in life-long learning in their profession and acquire professional engineering registration.
27
PEO1: Succeed and excel in developing sound solutions to civil engineering problems.
0
PEO3; Work effectively and conduct themselves ethically in their professional environment and grow in their careers working on projects designed for the well being of their society.
33
PEO2: Communicate and work competently in one or more core of civil engineering areas of practice or through graduate studies.
50(*)
Student Satisfaction Survey
70(*)
Performance of the Head of the Academic Department Survey
25
25
25
25
25
25
25
25
25
30
30
30
25
25
25
25

Fig. 8. GRL Model at the Program Level.

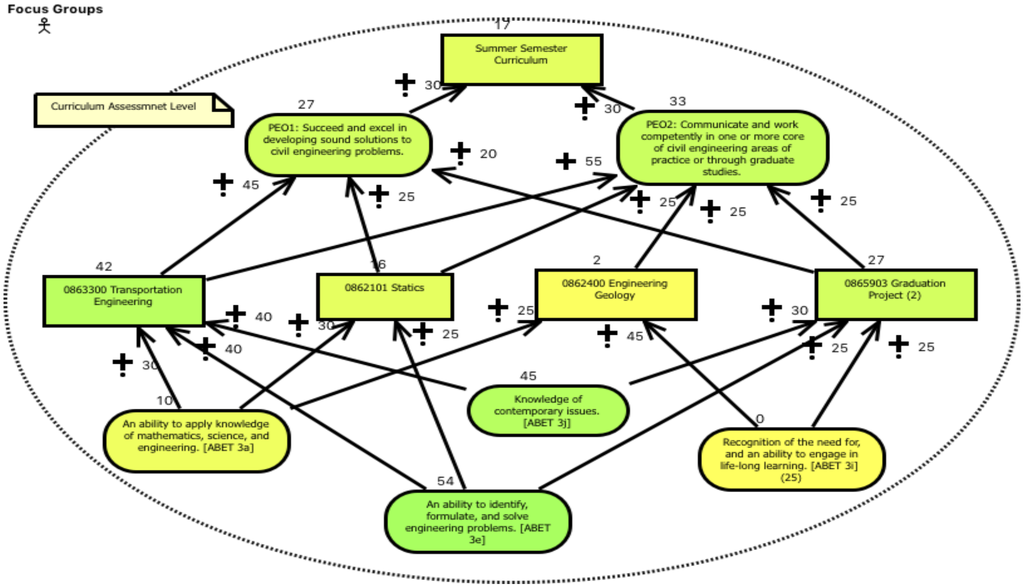

Focus Groups
Curriculum Assessmnet Level
17
Summer Semester Curriculum
27
PEO1: Succeed and excel in developing sound solutions to civil engineering problems.
33
PEO2: Communicate and work competently in one or more core of civil engineering areas of practice or through graduate studies.
42
0863300 Transportation Engineering
16
0862101 Statics
2
0862400 Engineering Geology
27
0865903 Graduation Project (2)
10
An ability to apply knowledge of mathematics, science, and engineering. [ABET 3a]
45
Knowledge of contemporary issues. [ABET 3j]
0
Recognition of the need for, and an ability to engage in life-long learning. [ABET 3i] (25)
54
An ability to identify, formulate, and solve engineering problems. [ABET 3e]
30
30
20
55
45
25
25
25
25
40
30
25
25
30
40
30
45
25
25

Fig. 9. GRL Model at the Curriculum Level.

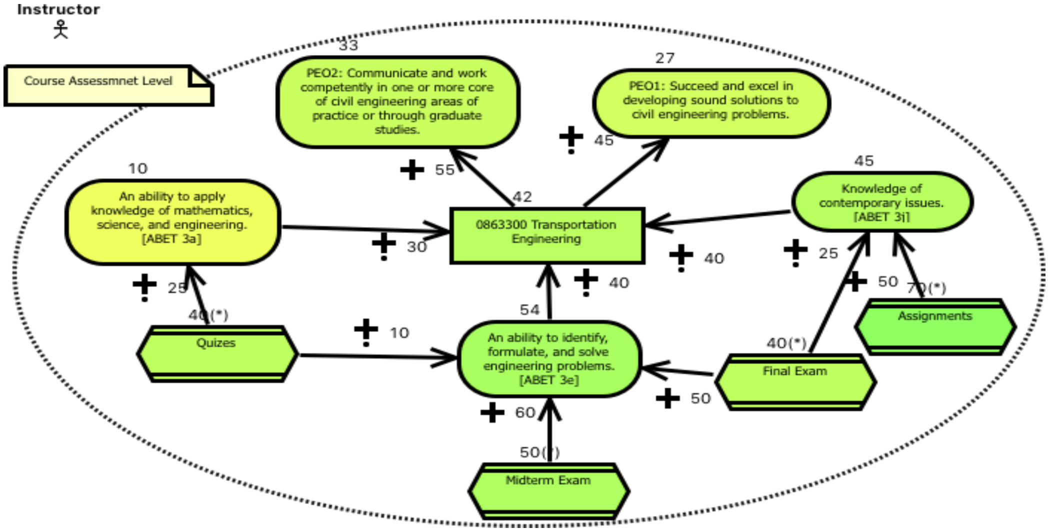

Instructor
Course Assessmnet Level
33
PEO2: Communicate and work competently in one or more core of civil engineering areas of practice or through graduate studies.
27
PEO1: Succeed and excel in developing sound solutions to civil engineering problems.
10
An ability to apply knowledge of mathematics, science, and engineering. [ABET 3a]
42
0863300 Transportation Engineering
45
Knowledge of contemporary issues. [ABET 3j]
54
An ability to identify, formulate, and solve engineering problems. [ABET 3e]
40(*)
Quizes
40(*)
Final Exam
70(*)
Assignments
50(*)
Midterm Exam
55
45
30
40
40
25
25
50
10
50
60

Fig. 10. GRL Model at the Course Level.

A. Evaluation of the Multi-level GRL Modelling Approach.

The proposed multi-level GRL modeling approach provides a model-driven technique in defining the strategic educational goals augmented with metric indicators. It is a rationale tool for modeling, analyzing, validating and documenting the learning assessment process that would be useful in detecting and anticipating any deficiencies in meeting learning goals and objectives where corrective actions can be made. Table II describe a sample of the output of the multi-level goal models. These results are the initial measures which will be used for future continues improvements.

Several challenges were addressed during the case study practice due to large number of participated constituencies and lack of quality former performance measures. In such modeling approach, teams from different disciplines are required to meet periodically to discuss the modeling structure, define the modeling elements and relationship between them and assign the measure values of the model elements, such as contribution values. This may increase the chance of error prone and increase the period of becoming familiar with modeling approach.

Several group decision approaches, and techniques have been discussed in [24] that can be used to ease this challenge. We used a Round-Table Discussion and Consensus (RTD&C) approach where teams assembled in a dialog setting. Groupings of related choices, contained in models, are put up on a screen, and the team members are asked to discuss the model elements and assign their relative weights to each choice in each grouping.

The other challenges were related to how much the measure of the model elements are accurate. It is obvious that the validity of multi-level GRL models depends on the accuracy of the model element measures. Though, we found based on our practice that the accuracy of GRL modeling results deviate towards the improvement as the time proceeds and the participated constituencies are familiar with approach.

During the practice in the case study, we found that the proposed approach can be used as goal modeling data mining, where patterns can be discovered in a large and complex related data. Through the technique of digging and aggregating in the multi-level GRL goals, we were able to detect any deficiencies at the low assessment levels, such as the course or curriculum levels that would affect the top assessment level.

TABLE. II. THE DEFINITION STEP OF THE CASE STUDY

Institution	Vision= 31	Mission= 31	
Faculty	Mission= 15		
Department	Mission= 25		
Program	PEOs= 32		
Curriculum	PEO1= 27	PEO2= 33	
Transportation Engineering Course	SO1= 10	SO2= 54	SO4= 45

VI. CONCLUSIONS AND FUTURE WORK

Strategic educational goals in educational organizations with multiple level of management is promoted as a sort of correlated and clustered data. The paper-based approaches often neglect the relationship and dependencies between the educational goals during the evaluation of strategic educational goals augmented with quantitative indicators. This would complicate their analysis, produces ambiguity of the results, and causes different interpretations by the management. We propose a multi-level goal modelling approach that provides a model-driven method to improve the assessment of strategic educational goals. The approach handles the group of data efficiently and allows the individual and group level assessment to take effect simultaneously. As a future work, we plan to include the service operation procedures (SOPs) in the goal modelling to study the effects of services such as admission, registration, withdraw, etc. on the strategic educational goals.

ACKNOWLEDGMENT

A special thanks for the ABET committee at the Faculty of Engineering of AAU.

REFERENCES

[1] M. Alhaj, "Towards model-based evaluation process of learning outcomes in academic institutions", 7th International Conference on Information and Education Technology (ICIET 2019), Aizu-Wakamatsu, Japan, March 29-31, 2019.

[2] ITU-T, "Recommendation z.151 (11/08): User requirements notation (URN)-language definition," Switzerland, 2008.

[3] A. Gelman, “Multilevel (hierarchical) modeling: what it can and cannot do”, American Statistical Association and the American Society for Quality TECHNOMETRICS, Vol. 48, No. 3, Auguat 2006.

[4] A. V. Diez-Roux, “Multilevel analysis in public health research”, annual review of public health, vol. 21, p.p.171-192, 2000.

[5] D. Kaplan, P. R. Elliott, “A Model-based approach to validating education indicators using multilevel structural equation modeling”, Journal of Educational and Behavioral Statistics, vol 22, issue 3, 1997.

[6] jUCMNav,2017, http://jucmnav.softwareengineering.ca/ foswiki/Projet SEG

[7] J. Heo, D. Ham, S. Park, C. Song and W. C., Yoon, "A framework for evaluating the usability of mobile phones based on multi-level”, hierarchical model of usability factors," in Interacting with Computers, vol. 21, no. 4, pp. 263-275, Aug. 2009.

[8] W. H. Sanders, "Integrated frameworks for multi-level and multi-formalism modeling." Proceedings 8th International Workshop on Petri Nets and Performance Models (Cat. No.PR00331) (1999): 2-9.

[9] M. Comuzzi, C. Kotsokalis, C. Rathfelder, W. Theilmann, U. Winkler, G. Zacco, "A Framework for multi-level sla management", Service-Oriented Computing. ICSOC/ServiceWave 2009 Workshops. ServiceWave 2009, ICSOC 2009. Lecture Notes in Computer Science, vol 6275. Springer, Berlin, Heidelberg.

[10] J. B. Yang and P. Sen, "A general multi-level evaluation process for hybrid MADM with uncertainty," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 24, no. 10, pp. 1458-1473, Oct. 1994.

[11] R. Tawhid et. al, “Towards outcome-based regulatory compliance in aviation security”, 20th IEEE International Requirements Engineering Conference (RE), p.p. 267-272, 2012.

[12] M. Alhaj, K. Mallur., B. Stepien. and L. Peyton, “Towards a model-based approach for developing and QA of online business processes”, 8th International Conference on Information and Communication Systems (ICICS), IEEE, 2017.

[13] S. Ghanavati, D. Amyot, L. Peyton, “Compliance analysis based on a goal-oriented requirement language evaluation methodology”, 17th IEEE International Requirements Engineering Conference, IEEE, 2009.

[14] L. Suskie, “Assessing student learning: a common sense guide”, 3rd Edition, ISBN: 978-1-119-42693-6, Feb 2018.

[15] A. Gastli, A. Al-Habsi, D. Al-Abri, “Innovative program and course outcomes' assessment tools”, 39th IEEE Frontiers in Education Conference, IEEE, USA, 2009.

[16] R. Delyser, M. A. Hamstad, “Outcomes based assessment and a successful ABET 2000 accreditation at the University of Denver”, FIE '00 Proceedings of the 30th Annual Frontiers in Education, vol. 01, USA, 2000.

[17] Y. Kwok-Bun, “Effective course-based learning outcome assessment for ABET accreditation of computing programs”, Journal of Computing Sciences in Colleges, Volume 22 Issue 4, April 2007.

[18] M. Besterfield-Sacre et. al, “Defining the outcomes: a framework for EC-2000”, IEEE Transactions on Education, Volume: 43 , Issue: 2 , May 2000.

[19] Y. M. Reddy, H. G. Andrade, "A review of rubric use in higher education", Semanticscholar, 2010.

[20] A. Gastli, A. Alhabsi, D. Al-Abri, "Innovative program and course outcomes' assessment tools", Frontiers in Education Conference, IEEE, 2009.

[21] Al-Ahliyya Amman University, https://www.ammanu.edu.jo/ENglish/HomeP/Home.aspx

[22] E. J. Caruana, M. Roman, J. Hernández-Sánchez, P. Soll, “Longitudinal studies”, Journal of Thoracic Disease, December 2015.

[23] ABET, 2019, http://www.abet.org/accreditation/accreditation-criteria/

[24] O. Akhigbe, “Creating quantitative goal models”, Governmental Experience, International Conference on Conceptual Modeling, pp 466-473, 2014.

A Modified Weight Optimization for Artificial Higher Order Neural Networks in Physical Time Series

Noor Aida Husaini[1], Rozaida Ghazali[2]
Nureize Arbaiy[3], Norhamreeza Abdul Hamid[4]
Faculty of Computer Science & Information Technology
Universiti Tun Hussein Onn Malaysia, Johor, Malaysia

Lokman Hakim Ismail[5]
Faculty of Civil Engineering and Built Environment
Universiti Tun Hussein Onn Malaysia
Johor, Malaysia

***Abstract*—Many methods and approaches have been proposed for analyzing and forecasting time series data. There are different Neural Network (NN) variations for specific tasks (e.g., Deep Learning, Recurrent Neural Networks, etc.). Time series forecasting are a crucial component of many important applications, from stock markets to energy load forecasts. Recently, Swarm Intelligence (SI) techniques including Cuckoo Search (CS) have been established as one of the most practical approaches in optimizing parameters for time series forecasting. Several modifications to the CS have been made, including Modified Cuckoo Search (MCS) that adjusts the parameters of the current CS, to improve algorithmic convergence rates. Therefore, motivated by the advantages of these MCSs, we use the enhanced MCS known as the Modified Cuckoo Search-Markov Chain Monté Carlo (MCS-MCMC) learning algorithm for weight optimization in Higher Order Neural Networks (HONN) models. The Lévy flight function in the MCS is replaced with Markov Chain Monté Carlo (MCMC) since it can reduce the complexity in generating the objective function. In order to prove that the MCS-MCMC is suitable for forecasting, its performance was compared with the standard Multilayer Perceptron (MLP), standard Pi-Sigma Neural Network (PSNN), Pi-Sigma Neural Network-Modified Cuckoo Search (PSNN-MCS), Pi-Sigma Neural Network-Markov Chain Monté Carlo (PSNN-MCMC), standard Functional Link Neural Network (FLNN), Functional Link Neural Network-Modified Cuckoo Search (FLNN-MCS) and Functional Link Neural Network-Markov Chain Monté Carlo (FLNN-MCMC) on various physical time series and benchmark dataset in terms of accuracy. The simulation results prove that the HONN-based model combined with the MCS-MCMC learning algorithm outperforms the accuracy in the range of 0.007% to 0.079% for three (3) physical time series datasets.**

Keywords—Modified Cuckoo Search-Markov Chain Monté Carlo; MCS-MCMC; neural networks; higher order; time series forecasting

I. Introduction

Time series forecasting involves developing a model or method that captures or describes the observed time series in order to understand the underlying causes. This research field looks for the “why” behind the time series dataset. This often involves making assumptions about data forms and breaking down time series into constitutional components [1, 2]. The challenge in time series forecasting is to provide a selection of techniques to better understand a dataset. In order to understand the past and predict the future event, it is important to analyze and optimize time series data using appropriate algorithms to understand underlying causes. There are many types of time series. For example; physical, financial and so forth [1, 3-5]. Time series forecasting have been addressed using classic methods such as the Autoregressive Integrated Moving Average (ARIMA) [6, 7], the Autoregressive Moving Average (ARMA) [7] and more. This linear model is the perfect choice for modeling time series events. However, they did not produce satisfactory results because they assumed a linear relationship between the past values of the series and ignored the non-linear relationships between these models.

Contrary, non-linear model such as Neural Networks (NN) has shown better performance as compared to linear models. Not to mention, it has been applied in dealing with issues of time series forecasting [8-12]. The NN is a type of parallel computer structure, which several of processing units are linked together thus that the computer’s memory is distributed, and information is passed in a parallel manner. Many NN architectures and algorithms have been developed thus far, namely multilayer feedforward networks, deep learning methods and so on [12-14]. Of these networks, the interest is gradually shifting towards using feedforward networks. Multilayer Perceptron (MLP), a class of feedforward networks, has been found to perform best in broader applications related to forecasting issues [1, 8-11]. The MLP is well-known for having the ability to map both linear and non-linear relationship if the number of nodes and layers are given sufficiently. However, MLP needs excessive learning time which may lead to overfitting [15, 16]. This is more likely to happen to the networks with many processing units and results in poor generalizability. The ability to generalize, that is to produce outputs from unknown inputs, is critical when the NN is used in time series forecasting. For this reason, networks with few parameters are preferred, fair enough to provide an adequate fit in order to avoid over-training [2, 15].

Therefore, to correct this failing, some Higher Order Neural Networks (HONN) is suggested. In this study, two (2) types of HONN were highlighted; Pi-Sigma Neural Network (PSNN) [17] and Functional Link Neural Network (FLNN) [18]. The PSNN utilizes product units at the output units that indirectly incorporate the capabilities of HONN while using a fewer number of weights and processing units. It has a regular structure, exhibits much faster learning, and is open to the incremental addition of units to attain a desired level of complexity. Meanwhile, the FLNN removes the need for hidden layers and hidden nodes by utilizing a higher order term

This work was funded by the Research Management Centre, Universiti Tun Hussein Onn Malaysia (Research Fund E15501).

to expand its input spaces into higher dimensional space within the single layer units. This simple architecture reduced the number of trainable parameters needed whilst reduces the learning complexity during the network training [19]. Taken as a whole, HONN are simple in their architecture and have fewer number of trainable parameters to deliver the input-output mappings as compared to the standard NN.

The standard method to train the NN is the well-known Backpropagation (BP) algorithm [20]. The existing BP algorithm, however, has several limitations including easily stuck into local minima, especially when dealing with highly non-linear problems [15]. The BP algorithm is also very dependent on the choices of initial values of the weights as well as other parameters. For instance, the BP algorithm is generally very slow as it requires small learning rates for stable learning. The momentum variation is usually faster than straightforward gradient descent since it allows higher learning rates while maintaining stability. However, it is still too slow for many practical applications.

Therefore, we used the Modified Cuckoo Search-Markov Chain Monté Carlo (MCS-MCMC) learning algorithm [21], that employs the learning rules to find the optimal weights in HONN models, thus overcome the BP drawbacks for this forecasting issue, and apply this method to several physical time series datasets. The results were compared with standard MLP and several HONN-based models. This MCS-MCMC used to enhanced the Modified Cuckoo Search (MCS) [22] by adopting Markov Chain Monté Carlo (MCMC) random walk. Those can be achieved by Markov chain mixing and integrated autocorrelation of a function of interest [23]. Therefore, it is useful in speeding up the convergence rate and obtaining higher accuracy rate.

Following this section, this paper is organized as follows: Section II presents the Related Works, followed by Section III which discuss the Architecture of HONN. Section IV poses the Experimental Results and Section V examines the Computational Results. Finally, Section VI concludes the work done.

II. Related Works

Weight optimizations are made in a wide range of diverse disciplines. Some methods that can be used to update weights in NN are BP, Genetic Algorithm (GA) [24, 25], Support Vector Machine (SVM) [26] and more. The concept of weight optimization by NN has become an active research field. It goes without saying that Swarm Intelligence (SI) played a role too. Among those swarm-based algorithms that have achieved significant popularity in the last few years are Evolutionary Algorithm (EA) [27, 28], Differential Evolution (DE) [29, 30], Artificial Bee Colony [16, 31] and Cuckoo Search [32].

The work presented in [33] combines Particle Swarm Optimization (PSO) and Extreme Learning Machine (ELM) to forecast the inflation rate in Indonesia. It uses PSO to optimize weight in order to obtain the optimal input values in ELM. In [34], the work binds the Ant Colony Optimization (ACO), PSO and 3-Opt algorithms. The PSO algorithm is used to optimize the parameter values used in the ACO algorithm for city selection operations, and defines the significance of inter-city pheromone and distances. 3-Opt heuristic approach to boost the local solutions is applied to the proposed method. The performance of the combined method becomes very significant in terms of solution quality and robustness. In the meantime, the research in [35] dealt with Whale Optimization Algorithm to optimize the weights and biases. Based on the findings, this algorithm has demonstrated the ability to solve a wide range of optimization issues and surpass the BP algorithm.

In conjunction with that, [36] presented GA with DE to change the weight parameters encoded within the structure by optimizing the network topology using GA and set the network weights using DE. Similar to [36], [37] combines GA and NN to increase the NN performance in diagnosing coronary artery disease. This somewhat shows surprising results which make the levels of accuracy, sensitivity and specificity achieved by that combination. In another study, [38] optimized the weight to speed up the convergence rate by reparametrizing the weight vectors in NN. Weight optimization is also studied by [39] using PSO. In his work, he combined the multiresolution analysis techniques with NN to forecast the next-day event. The findings suggested both results and good forecasting efficiency. Other research conducted by [40] used grid search technique to calculate the best value of SVM parameters. The use of those technique is crucial to forecast the time series event. The result shows that the SVM outperformed NN in terms of Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

In particular, the key reason why weight parameters are optimized is to prevent local minima and convergence speed. This is because, weights are the relative strength of node-to-node connections in NN. Besides, those optimization treats important topics such as having a particular way of manipulating and expanding the problem’s search space, which provides a detailed overview of how to manage such continuous domains. Instead, one of the well-known solutions is to find values of the variables that optimize the objectives. However, the variables are always limited, or somehow constrained. Therefore, in order to identify those values, experiments should focus on optimizing the objective functions or error functions due to the use of a common randomization arbitration and local search. Those parameter needs to be optimized subsequently to build up such appropriate and effective models. Once the effective models being developed, then the parameter is in its optimality conditions. It is however, the need for thorough research in order to evaluate the correct parameter measurement is still in doubt.

III. Architecture of HONN

In this study, the MCS-MCMC learning algorithm [21] is used to search for optimal weight parameters than can minimize the objective function in PSNN and FLNN network models. We replaced BP algorithm in the standard PSNN and FLNN with MCS-MCMC learning algorithm. The replacement is made to overcome the gradient-based learning algorithm drawbacks in BP algorithm that are slow, and easily get stuck into local minima [15]. Table I indicates the needs of MCS-MCMC that overcome the existing BP and MCS learning algorithm.

TABLE. I. COMPARISON OF BP, MCS AND MCS-MCMC LEARNING ALGORITHM

BP	MCS	MCS-MCMC
• Stuck into local minima. • Very dependent on the initial weights. • Need more parameter to be set up.	• Caters slow convergence encountered by BP. • Less parameter to be set up.	• Reduce complexity. • Speed up convergence rate. • Initialize weight value for better way/solutions and abondoned poor values.

According to Table I, the MCS-MCMC is used for weight initialization and weight update (replacing the BP algorithm in the standard PSNN and FLNN). The weights and biases were calculated and updated for the complete training that represents the architecture. Those can be achieved by starting it with random values followed by several repeated attempts on discovering better solutions and abandoning the poor values. The architecture of Pi-Sigma Neural Network-Markov Chain Monté Carlo (PSNN-MCMC) and Functional Link Neural Network-Markov Chain Monté Carlo (FLNN-MCMC) are presented in Fig. 1 and Fig. 2.

$x_1, x_2, \ldots x_n$ denotes input vectors, w_{ij} denotes adjustable weights for input vectors to linear summing unit, σ is the non-linear activation function, $h_1, h_2, \ldots h_l$ indicates the summing units, y is the output node and w_{jk} is the fixed weights from linear summing units to the output layer. Step-by-step process in PSNN-MCMC:

Step 1: Initialize weights w_{ij} from input vector to the linear summing unit h_l with a random number using MCS-MCMC learning algorithm. Those random weights are evaluated from layer-to-layer to improve the searching strategies to get the optimal weights set.

Step 2: Transform the optimization parameters (weights and biases) into the objective function.

Step 3: Feed the objective function into the MCS-MCMC learning algorithm to search for optimal weight parameters.

Step 4: Calculate error.

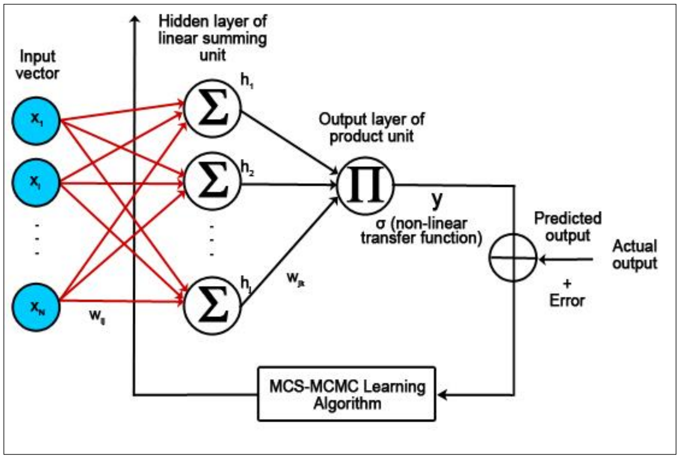

Input vector
Hidden layer of linear summing unit
x1
xi
xN
wij
h1
h2
hl
Σ
Σ
Σ
Output layer of product unit
Π
y
σ (non-linear transfer function)
wjk
Predicted output
Actual output
+
Error
MCS-MCMC Learning Algorithm

Fig. 1. The Architecture of PSNN-MCMC.

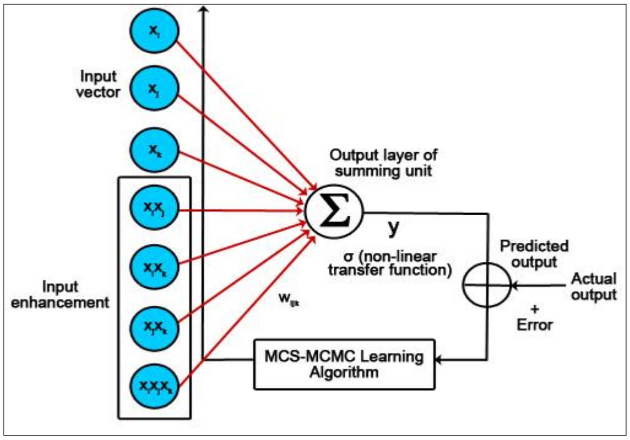


Fig. 2. The Architecture of FLNN-MCMC.

x_i, x_j, x_k is the input vector, w_{ijk} is the adjustable weight, y is the output, and σ is the non-linear activation function.

Step-by-step process in FLNN-MCMC:

Step 1: Initialize weights w_{ijk} with a random number using MCS-MCMC learning algorithm.

Step 2: In the initial process, transform the standard FLNN architecture (weight and biases) into the objective function.

Step 3: Feed the objective function, along with the training data, into the MCS-MCMC learning algorithm to search for optimal weight parameters to minimize the objective function.

Step 4: Tune the weight changes using the MCS-MCMC learning algorithm based on the error calculation (the difference between actual and predicted outputs).

Step 5: Obtain the optimal weights set from the training phase and used upon unseen data for forecasting.

IV. EXPERIMENTAL RESULTS

A. Data Preparation

Appropriate datasets should be provided to determine the problems encountered and evaluate the performance of the proposed PSNN-MCMC and FLNN-MCMC, and other models; standard PSNN, Pi-Sigma Neural Network-Modified Cuckoo Search (PSNN-MCS), standard FLNN, Functional Link Neural Network-Modified Cuckoo Search (FLNN-MCS), and standard MLP. The performance are evaluated based on the lowest Mean Squared Error (MSE) [41, 42] and Root Mean Squared Error (RMSE) [43]. Based on the previous records, the maximum, minimum and average measurements of three (3) datasets are tabulated in Table II.

TABLE. II. THE DATASETS EVALUATIONS

Dataset	Minimum	Maximum	Average	Data Size
Relative Humidity	69.5000	98.1000	85.9035	50, 840
Temperature	23.7000	29.5000	26.7543	1, 813
Santa Fe Laser	0	255	59.8661	3, 972

Relative Humidity: The datasets were collected from Malaysian Meteorological Department (MMD). Each dataset consists of 50, 840 instances which are covered from year of 1992 until 2009 [44].

Temperature: The datasets were collected from MMD. Each dataset consists of 1, 813 that covers over year of 1992 until 2009 [44].

Santa Fe Laser: A univariate time series derived from laser-generated data recorded from a Far-Infrared-Laser in a chaotic state. This benchmark datasets are composed of a clean low-dimensional non-linear and stationary time series with the total number of 3, 972 instances.

The reason for choosing these datasets are due to the stability they owned compared to other datasets. The stability is depending on the types of data and factors affecting them [45, 46]. For instance, the time series signals were observed on a highly non-stationary and/or non-linear range [47, 48]. Non-stationary is a common property to vary time-series models, which means, a variable has no clear tendency to return to a constant value or a linear trend. To note, the stability is the key to predictability. Therefore, a stable dataset is needed to predict the current trend. These physical time series data, later, were fed to all NN to capture the underlying rules of the movement.

B. Data Pre-processing

Mostly, data gathering somehow are loosely controlled. Thus, resulting in outliers, impossible data combinations, and may contains missing values. Therefore, the data need to be pre-processed to avoid errors and misleading results Fig. 3. The data pre-processing involves cleaning, shifting and normalizing the raw data into a format that improves the performance of the subsequent modules [18, 49].

C. Data Partition

Data partitioning is highly required by NN to obtain best NN models. Hence, in this study, we divide the datasets into three (3) partitions: 60% for training, while 20% for both testing and validation.

Training Set: Served the model for training purposes which allows the model to produce an output closer to the target value. Therefore, it must have more significant portion than the data being used for testing and validation.

Validation Set: Used to evaluate a given model, in which the sample of data used to provide an unbiased evaluation of a model fit on the training dataset fine-tunes the model. This set is also essential to avoid overfitting.

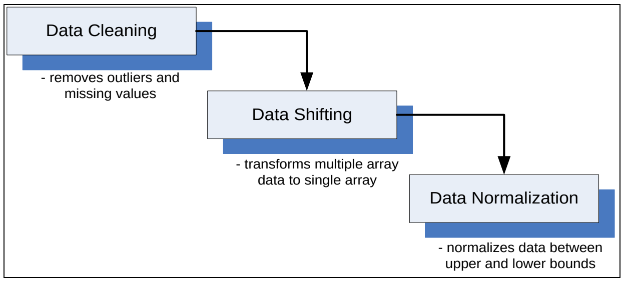

Data Cleaning
- removes outliers and missing values
Data Shifting
- transforms multiple array data to single array
Data Normalization
- normalizes data between upper and lower bounds

Fig. 3. Data Pre-Processing Process.

Testing Set: Describes how the models will perform on new, unseen data in order to evaluate the model. This sample provides an unbiased evaluation of a final model fit on the training dataset. It is only used once a model is thoroughly trained.

The split ratio of the datasets mainly relies on two (2) criteria. First, the total number of samples in the dataset. Second, the actual model going to be trained. Some models need substantial data to train upon. Therefore, in this study, more massive training sets should be optimized. Models with very few hyperparameters (e.g., momentum, learning rate, etc.) will be easy to validate and tune. As is, the validation set can probably reduce. However, if the model has many hyperparameters, an extensive validation must be set as well. All in all, like many other things in NN, the training-testing-validation split ratio is also quite specific based on some instances, and it gets easier to make a judgment as more training used.

D. Parameters Settings

The parameters of an NN are learned during the training stage. Learning (or training) is a process by which the tunable weights of a network are adapted through a continuous process of simulation whereas the network is embedded. The most basic method of training a network is a trial-and-error procedure [15]. During the learning phase, the network learns until its weight continues to tweak. The same set of data is then processed many times as the connection weight continues to improve. Parameters must be specified during training for any given NN architecture. For all network models, input nodes are set between 5 and 7 nodes, higher nodes / nodes between 2 and 5 (except for standard MLP) and one (1) for output nodes. The parameter settings for all network models are tabulated in Table III.

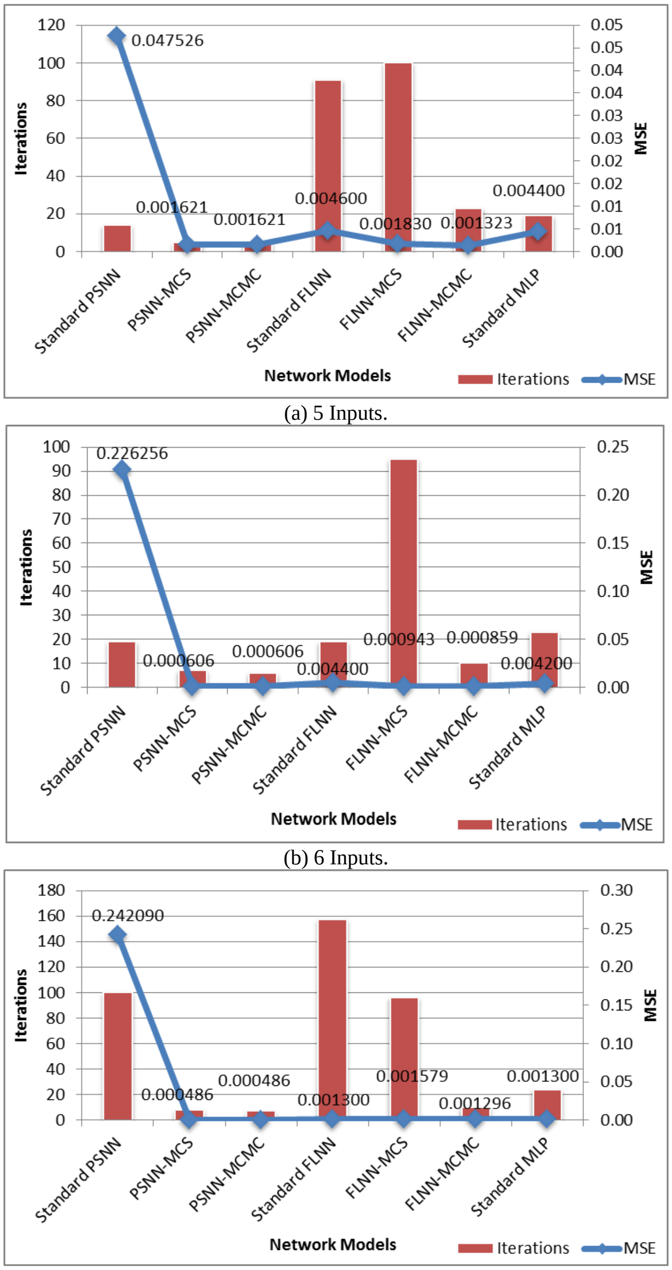
TABLE. III. PARAMETER SETTINGS FOR ALL NETWORK MODELS

Parameters	**Values**	**References**
Initial weights	$[0.25, 0.75]$	[15]
Learning Rate	0.2	[15]
Momentum	0.3	[15]
Minimum Error	0.001	[15]
Epoch	1000	[15]
Initial Value, A	1	[23]
Step size, α	0.01	[23]
Probability, P_{α}	0.25	[22]
Initial Value, θ	1	[23]
n	5	[23]
a	4	[23]
Minimum Error	0.001	[15]
Number of Generation	1000	[22]
Input Nodes	5 to 7	[15]
Network's Order	2 to 5 (for rest of NN models) 3 to 8 (for MLP)	[15]
Output Node	1	[15]
Transfer Function	Sigmoid	[15]

V. Computational Results

A. Relative Humidity Dataset

Referring to Fig. 4, the MSE results for Relative Humidity with 5 to 7 input nodes are visualized. As the 5 inputs were supplied, FLNN-MCMC, PSNN-MCMC and PSNN-MCS lead the ranks. When inputs 6 and 7 were loaded, PSNN-MCMC, PSNN-MCS and FLNN-MCMC outperformed. Seemingly, based on the results, the performances of the network in which the learning method had been replaced by MCS-MCMC learning algorithm are much preferable compared to the networks with standard MCS algorithm.



(a) 5 Inputs.

(b) 6 Inputs.

(c) 7 Inputs.

Fig. 4. Performance Comparison on Relative Humidity.

B. Temperature Dataset

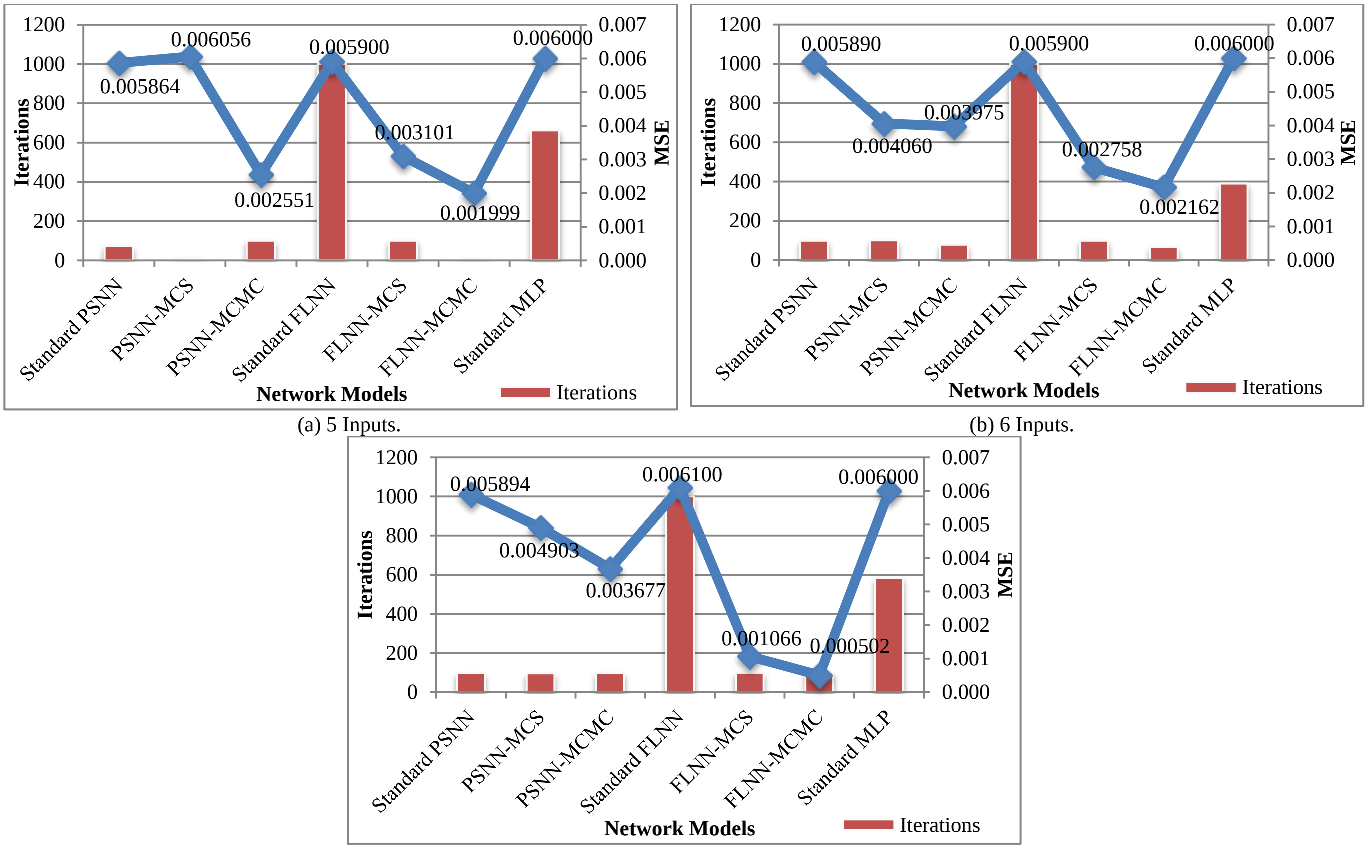
Fig. 5 graphically shows the performance comparison for all the networks on Temperature dataset. According to the results plotted in Fig. 5, the first, second and third ranks are FLNN-MCMC, PSNN-MCMC and FLNN-MCS for 5 inputs, FLNN-MCMC, FLNN-MCS and PSNN-MCMC for 6 and 7 inputs. From these results, it is said that the incorporation of MCS-MCMC learning algorithm into both PSNN and FLNN network models could help to minimize the error rate, thus assists the network to converge quickly. As it has been pointed out, FLNN-MCMC shows the least MSEs compared to all network models generated. Therefore, by having the least MSE, it combines both the estimator's variance and its bias to the extent that the estimated value is derived from the truth. In addition, the positive tendency in the Temperature dataset itself indicates that the data have a strong influence / fluctuation that is stable enough to handle the network model integrated with the MCS-MCMC learning algorithm.

C. Santa Fe Laser Dataset

In view of inputs 5, 6 and 7, the FLNN-MCMC also outperformed the other network models for 60:20:20 data partition. Fig. 6 shows the results with respect to iterations and MSE values. From these statistics, it can be noted that the FLNN-MCMC network model performed better than the other network models with stable results even when dealing with the Santa Fe Laser dataset's temporal behavior.

The current study includes trials of MCS-MCMC learning algorithm on various network models. From the results, it is proved that, in this study, it is affirmative that the networks with MCS-MCMC learning algorithm were well generalized and showed least error compared to other network models, which could represent non-linear function. The MCS-MCMC's existence as the learning algorithm that replaces the existing BP algorithm enabled fast and rapid training. A significant advantage of the MCS-MCMC is that the learning algorithm can automatically adjust better parameters to find excellent parameter values with little user interference, which being accomplished through Markov chain mixing and a functional of interest integrated autocorrelation. Overall, the use of MCS-MCMC learning algorithm was discovered to be able to perform on various ranges of datasets.

The MCS-MCMC is developed for initializing and updating the weights in HONN-based models. The use of Swarm Intelligence (SI) techniques in MCS-MCMC allows it to expand their input space to a higher dimensional space where linearity separable is possible has led to a significant effect on improving the network performance. The network is computationally efficient and is capable of modelling non-linear input-output mappings when learning the time series data, thus justified the potential use of this model by practitioners. Besides, the results clearly showed that the MCS-MCMC substantially at par with the computational efficiency of the training process, and has been developed in order to produce more realistic and acceptable results.

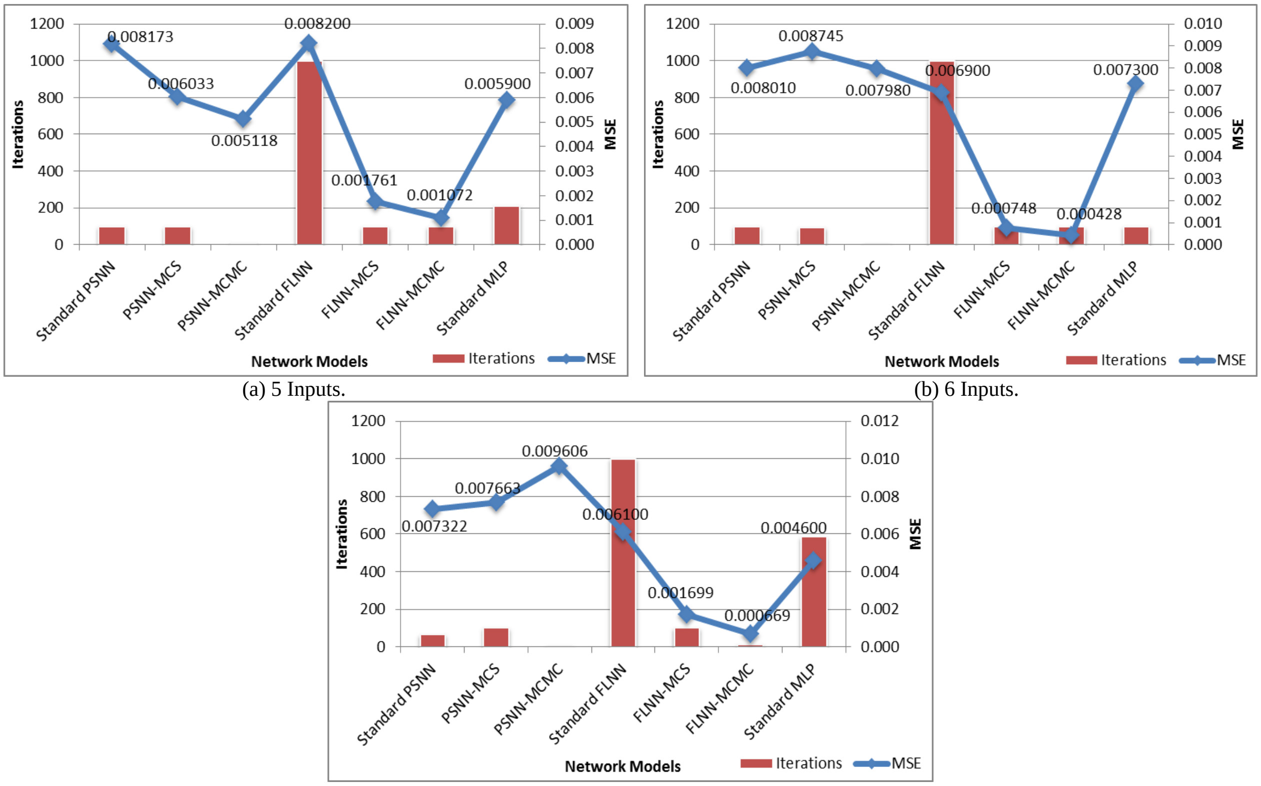


(a) 5 Inputs.

(b) 6 Inputs.

(c) 7 Inputs.

Fig. 5. Performance Comparison on Temperature.



(a) 5 Inputs.

(b) 6 Inputs.

(c) 7 Inputs.

Fig. 6. Performance Comparison on Santa Fe Laser.

D. Discussions

In this section, several issues raised by different NN comparisons are addressed. Because the results presented previously include extensive simulations, this section describes the observations obtained from the entire experimental results.

1) Model performances based on ranking: The simulation results in Section V, Subsection A were summarized in Tables IV to VI. This tables cover inputs ranges from 5 to 7 and seven (7) network models. Table IV shows the overall rank for Relative Humidity on all networks.

From Table IV, the PSNN-MCMC outperformed other network models by getting the highest average ranking. This demonstrates that the accuracy rate is enhanced by integrating the MCS-MCMC learning algorithm with HONN. Table V indicates the overall rank for Temperature on all networks.

According to Table V, FLNN-MCMC outperformed the other network models by having the highest average rank. This is followed by FLNN-MCS and PSNN-MCMC in the second and third rank, respectively. Basically, those swarm-based learning algorithm helps to overcome the drawbacks of the existing BP algorithm. Table VI summarizes data on all networks from the Santa Fe Laser dataset.

The results in Table VI show that the FLNN-MCMC provides a lower MSE than the other network models. This is accompanied by FLNN-MCS that falls into the second place and standard MLP in the third place. Based on these outcomes, it is concluded that implementing the swarm-based learning algorithm in HONN helps network models converge with lower iterations and lower error rate. Therefore, improves the network performance indirectly.

TABLE. IV. OVERALL RANK FOR RELATIVE HUMIDITY ON ALL NETWORKS

Inputs	**Standard PSNN**	**PSNN-MCS**	**PSNN-MCMC**	**Standard FLNN**	**FLNN-MCS**	**FLNN-MCMC**	**Standard MLP**
5	7	3	2	6	4	1	5
6	7	2	1	6	4	3	5
7	7	2	1	5	6	3	4
Mean Rank	7.00	2.33	1.33	5.67	4.67	2.33	4.67
Overall Rank	7	2	1	6	4	2	4

TABLE. V. OVERALL RANK FOR TEMPERATURE ON ALL NETWORKS

Inputs	**Standard PSNN**	**PSNN-MCS**	**PSNN-MCMC**	**Standard FLNN**	**FLNN-MCS**	**FLNN-MCMC**	**Standard MLP**
5	4	7	2	5	3	1	6
6	5	4	3	6	2	1	7
7	5	4	3	7	2	1	6
Mean Rank	4.67	5.00	2.67	6.00	2.33	1.00	6.33
Overall Rank	4	5	3	6	2	1	7

TABLE. VI. OVERALL RANK FOR SANTA FE LASER ON ALL NETWORKS

Inputs	**Standard PSNN**	**PSNN-MCS**	**PSNN-MCMC**	**Standard FLNN**	**FLNN-MCS**	**FLNN-MCMC**	**Standard MLP**
5	6	5	3	7	2	1	4
6	6	7	5	3	2	1	4
7	5	6	7	4	2	1	3
Mean Rank	5.67	6.00	5.00	4.67	2.00	1.00	3.67
Overall Rank	6	7	5	4	2	1	3

2) The accuracy: In this section, we presented the result based on the percentage of RMSE and Accuracy. The RMSE used to measures how much error there is between the actual and the target output [42]. In other words, it tells how concentrated the data is around the line of best fit. In general, if the value of RMSE getting lower, the better performance will be produced.

Tables VII to IX show the experimental results on all datasets. The table consisted of six (6) elements. The first element indicates the network model; and the second element designates the best network structure. This is accomplished by the method of trial-and-error procedure [15]. The third element specifies the number of trainable weights. Those values are collected during experiments. The fourth element is the RMSE value acquired through Equation (1):

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}\left(P_i-\tilde{P}_i\right)^2}{n}} \qquad (1)$$

where n is the total number of data patterns, P_i and $\tilde{P}_i$ represent the actual and predicted output value, respectively. Equation (2) provides the sixth element (Accuracy in percentage). The simulation results later being compared in the form of accuracy rate.

$$Accuracy=\left(1-\frac{MSE}{2}\right)\times 100 \qquad (2)$$

where MSE is mean squared error [42].

TABLE. VII. EXPERIMENTAL RESULTS ON RELATIVE HUMIDITY

Network Model	**Best Network Structure**	**No. of Trainable Weights**	**RMSE**	**Accuracy (%)**
Standard PSNN	5-2-1	10	0.21801	97.624
PSNN-MCS	7-5-1	35	0.02205	99.976
PSNN-MCMC	7-2-1	14	0.02205	99.976
Standard FLNN	7-4-1	127	0.03606	99.935
FLNN-MCS	6-4-1	57	0.03071	99.953
FLNN-MCMC	6-4-1	57	0.02931	99.957
Standard MLP	7-7-1	56	0.03606	99.935

TABLE. VIII. EXPERIMENTAL RESULTS ON TEMPERATURE

Network Model	Best Network Structure	No. of Trainable Weights	RMSE	Accuracy (%)
Standard PSNN	5-3-1	15	0.07658	99.707
PSNN-MCS	6-2-1	12	0.06372	99.797
PSNN-MCMC	5-2-1	10	0.05051	99.872
Standard FLNN	5-4-1	31	0.07681	99.705
FLNN-MCS	6-4-1	57	0.05252	99.862
FLNN-MCMC	7-3-1	120	0.02241	99.975
Standard MLP	6-7-1	49	0.07746	99.700

TABLE. IX. EXPERIMENTAL RESULTS ON SANTA FE LASER

Network Model	Best Network Structure	No. of Trainable Weights	RMSE	Accuracy (%)
Standard PSNN	7-4-1	28	0.08557	99.634
PSNN-MCS	5-2-1	10	0.07767	99.698
PSNN-MCMC	5-3-1	15	0.07154	99.744
Standard FLNN	7-4-1	127	0.07810	99.695
FLNN-MCS	6-4-1	57	0.02735	99.963
FLNN-MCMC	6-4-1	57	0.02069	99.979
Standard MLP	7-8-1	64	0.06782	99.770

According to the results on Relative Humidity dataset (refer to Table VII), the HONN-based models being incorporated with MCS-MCMC learning algorithm give significant percentage around 99.953% to 99.976%. while for Temperature dataset (refer to Table VIII), the values vary from 99.797% to 99.975%. For Santa Fe Laser dataset (refer to Table IX), the FLNN-MCMC achieved highest percentage of Accuracy with the value of 99.979%.

The experimental results vary depending on the datasets. The algorithm can readily mapped the function if the data is sufficiently stable, thus delivering much better and stable outcomes. Otherwise, it could result in an extensive training algorithm. As the time series datasets exhibit a very strong trend, it shows obvious up and down movement. Therefore, during the training of such datasets, the networks were used to learn the precise values of each data point. This sometimes could lead the networks failed to respond well to the underlying chaotic structure within the data behaviour. Hence, to correctly predict the value from one point to another point is a challenging task.

3) Threat to validity and improvements: In this study, the fairness of experimentations involving SI technique are levelled to minimize threats to validity. One of major concerned was regarding the validity of parameter setting for each SI technique. In order to ensure fair comparisons, all parameter settings for all the network models, involving the input settings, learning rate and stopping criteria are set with the same value (revisit Section III, D). Another concerned was regarding the network structure for all the network models; the standard PSNN, PSNN-MCS, PSNN-MCMC, standard FLNN, FLNN-MCS, FLNN-MCMC and standard MLP. The network structure for those network models cannot be equivalent for all datasets in the experiment as they may yield unfair results. Therefore, to ensure fair prediction performance results, the network structure issue is addressed.

The critical part is on generalization. It is on how the network generates lowest MSE. For this reason, the best model is regarded to the NN structure that offers the greatest proportion of improvements. The simulation results are benchmarked against seven (7) NN models. The improvements for MCS-MCMC learning algorithm on both PSNN and FLNN for all datasets are measured in Equations (3) and (4). Let a be standard PSNN, b be PSNN-MCS, c be PSNN-MCMC, d be standard FLNN, e be FLNN-MCS, f be FLNN-MCMC and g be standard MLP.

$$Improvement_c(\%) = \frac{\left|\left(c - \frac{a+b+c+d+e+f+g}{7}\right)\right|}{c} \times 100\% \tag{3}$$

$$Improvement_f(\%) = \frac{\left|\left(f - \frac{a+b+c+d+e+f+g}{7}\right)\right|}{f} \times 100\% \tag{4}$$

$Improvement_c$ denotes improvement for PSNN-MCMC while $Improvement_f$ denotes improvement for FLNN-MCMC [42]. The overall improvements for PSNN-MCMC and FLNN-MCMC are tabulated in Tables X to XI. The findings on Table X show that the PSNN-MCMC provides significant improvement in all datasets where the PSNN-MCMC can improve the accuracy. This is also applicable to FLNN-MCMC in Table XI.

As can be seen from Tables X and XI, the MCS-MCMC learning algorithm can train and improve the accuracy of the HONN network model. Thus, it makes the best improvement on Relative Humidity dataset with the value of 0.707% on PSNN-MCMC and 0.670% on FLNN-MCMC when compared to other datasets. Both network models operate approximately 0.007 % to 0.079%.

TABLE. X. THE OVERALL IMPROVEMENTS OF PSNN-MCMC

Datasets	Network Structure	Improvement of PSNN-MCMC (%)
Relative Humidity	7-2-1	0.707
Temperature	5-2-1	0.116
Santa Fe Laser	5-3-1	0.079

TABLE. XI. THE OVERALL IMPROVEMENTS OF FLNN-MCMC

Datasets	Network Structure	Improvement of FLNN-MCMC (%)
Relative Humidity	6-4-1	0.670
Temperature	7-3-1	0.320
Santa Fe Laser	6-3-1	0.391

As the time series have chaotic behavior, this approach offers significant advantages over the standard network models such as improved simulations and lower error rate, due to their ability to better approximate complex, non-smooth and often discontinuous training datasets. To conclude, it is confirmed that HONN, when incorporated with MCS-MCMC learning algorithm, helps to overcome the drawback of the existing BP algorithm that prone to overfit and stuck into local minima. Thus, improve the network performance and increase the accuracy by getting the highest average ranking.

VI. Conclusion

The higher demands for SI techniques justify the need for a more effective, better solutions approach. The findings of this study will redound to the benefit of the SI field, considering that SI plays a vital role in optimization issues today. Therefore, the MCS-MCMC learning algorithm nailing down the optimal weight values in HONN which helped in dealing with slow convergence and poor generalization. Those are derived from the findings which later will be used to predict the time series event better. This study may also advantageous for certain sectors such as meteorological department that applies the non-linearity relationship in meteorological process. On the other hand, by obtaining outstanding performance on various ranges of time series datasets, it may reduce the risk in decision making. Thus, this approach significantly matches the idea. Therefore, the effectiveness of any decision depends upon the nature of a sequence of events preceding the decision. Furthermore, this study would be beneficial to the researchers, as it can provide baseline information on the different approach of SI and NN.

Acknowledgment

This work was funded by the Research Management Centre, Universiti Tun Hussein Onn Malaysia (Research Fund E15501).

References

[1] Gershenfeld, N.A. and A.S. Weigend, The future of time series: Learning and understanding, in Pattern Formation In The Physical And Biological Sciences. 2018, CRC Press. p. 349-429.

[2] Husaini, N.A., et al. Jordan pi-sigma neural network for temperature prediction. in International Conference on Ubiquitous Computing and Multimedia Applications. 2011. Springer.

[3] Costa, M., A.L. Goldberger, and C.-K. Peng, Multiscale entropy analysis of complex physiologic time series. Physical review letters, 2002. 89(6): p. 068102.

[4] Batt, R.D., S.R. Carpenter, and A.R. Ives, Extreme events in lake ecosystem time series. Limnology and Oceanography Letters, 2017. 2(3): p. 63-69.

[5] Bao, W., J. Yue, and Y. Rao, A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PloS one, 2017. 12(7): p. e0180944.

[6] Zhang, G.P., Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 2003. 50: p. 159-175.

[7] Said, S.E. and D.A. Dickey, Testing for unit roots in autoregressive-moving average models of unknown order. Biometrika, 1984. 71(3): p. 599-607.

[8] Bishop, C.M., Neural networks for pattern recognition. 1995: Oxford university press.

[9] Kolarik, T. and G. Rudorfer. Time series forecasting using neural networks. in ACM Sigapl Apl Quote Quad. 1994. ACM.

[10] Brath, A., A. Montanari, and E. Toth, Neural networks and non-parametric methods for improving real-time flood forecasting through conceptual hydrological models. Hydrology and Earth System Sciences Discussions, 2002. 6(4): p. 627-639.

[11] Shrestha, R.R., S. Theobald, and F. Nestmann, Simulation of flood flow in a river system using artificial neural networks. Hydrology and Earth System Sciences Discussions, 2005. 9(4): p. 313-321.

[12] Ali, Z., et al., Forecasting drought using multilayer perceptron artificial neural network model. Advances in Meteorology, 2017. 2017.

[13] Ryu, S., J. Noh, and H. Kim, Deep neural network based demand side short term load forecasting. Energies, 2017. 10(1): p. 3.

[14] Hewamalage, H., C. Bergmeir, and K. Bandara, Recurrent neural networks for time series forecasting: Current status and future directions. arXiv preprint arXiv:1909.00590, 2019.

[15] Ghazali, R., et al., The application of ridge polynomial neural network to multi-step ahead financial time series prediction. Neural Computing & Applications, 2008. 17: p. 311-323.

[16] Shah, H., et al., A quick gbest guided artificial bee colony algorithm for stock market prices prediction. Symmetry, 2018. 10(7): p. 292.

[17] Shin, Y. and J. Ghosh. The pi-sigma network: An efficient higher-order neural network for pattern classification and function approximation. in IJCNN-91-Seattle International Joint Conference on Neural Networks. 1991. IEEE.

[18] Giles, C.L. and T. Maxwell, Learning, invariance, and generalization in high-order neural networks. Applied optics, 1987. 26(23): p. 4972-4978.

[19] Garro, B.A., H. Sossa, and R.A. V´azquez. Design of artificial neural networks using differential evolution algorithm. in Proceedings of the 17th international conference on Neural information processing: models and applications. 2010. Springer-Verlag.

[20] Leung, H. and S. Haykin, The complex backpropagation algorithm. IEEE Transactions on Signal Processing, 1991. 39(9): p. 2101-2104.

[21] Husaini, N.A., R. Ghazali, and I.T.R. Yanto. Enhancing modified cuckoo search algorithm by using MCMC random walk. in 2016 2nd International Conference on Science in Information Technology (ICSITech). 2016. IEEE.

[22] Walton, S., et al., Modified cuckoo search: A new gradient free optimisation algorithm. Chaos, Solitons & Fractals, 2011. 44(9): p. 710-718.

[23] Hastings, W.K., Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 1970. 57(1): p. 97-109.

[24] Holland, J., Adaptation in natural and artificial systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. 1992, MIT Press: Ann Arbor, USA.

[25] Goldberg, D., Genetic algorithms in search, optimization and machine learning. 1989, Boston, USA: Addison Wesley.

[26] Vapnik, V.N., The nature of statistical learning. Theory, 1995.

[27] De Jong, K., Analysis of the behavior of a class of genetic adaptive systems. 1975, University of Michigan: Ann Arbor, MI.

[28] Fogel, L., A. Owens, and W. MJ, Artificial intelligence through simulated evolution. 1966, Chichester, UK: John Wiley.

[29] Storn, R. Differential evolution design of an IIR-filter. in IEEE International Conference on Evolutionary Computation. 1996. Nagoya.

[30] dos Santos Coelho, L. and D.L. de Andrade Bernert, An improved harmony search algorithm for synchronization of discrete-time chaotic systems. Chaos, Solitons & Fractals, 2009. 41(5): p. 2526-2532.

[31] Karaboga, D., B. Akay, and C. Ozturk. Artificial bee colony (abc) optimization algorithm for training feed-forward neural networks. in Proceedings of the 4th international conference on Modeling Decisions for Artificial Intelligence, ser. MDAI '07. 2007. Springer-Verlag.

[32] Yang, X.S. and S. Deb. Cuckoo search via Lévy flights. in Proceedings of the World Congress on Nature & Biologically Inspired Computing (NaBIC '09. 2009. India: IEEE Publications.

[33] Alauddin, M.W., W.F. Mahmudy, and A.L. Abadi, Extreme Learning Machine Weight Optimization using Particle Swarm Optimization to Identify Sugar Cane Disease. Journal of Information Technology and Computer Science, 2019. 4(2): p. 127-136.

[34] Gülcü, Ş., et al., A parallel cooperative hybrid method based on ant colony optimization and 3-Opt algorithm for solving traveling salesman problem. Soft Computing, 2018. 22(5): p. 1669-1685.

[35] Aljarah, I., H. Faris, and S. Mirjalili, Optimizing connection weights in neural networks using the whale optimization algorithm. Soft Computing, 2018. 22(1): p. 1-15.

[36] Mason, K., J. Duggan, and E. Howley. Neural network topology and weight optimization through neuro differential evolution. in Proceedings of the Genetic and Evolutionary Computation Conference Companion. 2017.

[37] Arabasadi, Z., et al., Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. Computer Methods and Programs in Biomedicine, 2017. 141: p. 19-26.

[38] Salimans, T. and D.P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. in Advances in neural information processing systems. 2016.

[39] Lahmiri, S., A variational mode decompoisition approach for analysis and forecasting of economic and financial time series. Expert Systems with Applications, 2016. 55: p. 268-273.

[40] Samsudin, R., A. Shabri, and P. Saad, A comparison of time series forecasting using support vector machine and artificial neural network model. Journal of applied sciences, 2010. 10(11): p. 950-958.

[41] Chae, Y.T., et al., Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings. Energy and Buildings, 2016. 111: p. 184-194.

[42] Hassim, Y.M.M. and R. Ghazali, Optimizing functional link neural network learning using modified bee colony on multi-class classifications, in Advances in Computer Science and its Applications. 2014, Springer. p. 153-159.

[43] Leva, S., et al., Analysis and validation of 24 hours ahead neural network forecasting of photovoltaic output power. Mathematics and computers in simulation, 2017. 131: p. 88-100.

[44] Department, M.M. Weather Forecast. 2010 [cited 2011 February, 18]; Available from: http://www.met.gov.my.

[45] Ribeiro, H.V., et al., Characterizing time series via complexity-entropy curves. Physical Review E, 2017. 95(6): p. 062106.

[46] Rounaghi, M.M. and F.N. Zadeh, Investigation of market efficiency and financial stability between S&P 500 and London stock exchange: Monthly and yearly forecasting of time series stock returns using ARMA model. Physica A: Statistical Mechanics and its Applications, 2016. 456: p. 10-21.

[47] Akram, U., et al., An Improved Pi-Sigma Neural Network with Error Feedback for Physical Time Series Prediction. International Journal of Advanced Trends in Computer Science and Engineering, 2019. 8: p. 276-284.

[48] Al-Jumeily, D., R. Ghazali, and A. Hussain, Predicting Physical Time Series Using Dynamic Ridge Polynomial Neural Networks. PLOS ONE, 2014. 9(8): p. e105766.

[49] García, S., J. Luengo, and F. Herrera, Data preprocessing in data mining. 2015: Springer.

Remote Sensing Satellite Image Clustering by Means of Messy Genetic Algorithm

Kohei Arai
Graduate School of Science and Engineering
Saga University, Saga City, Japan

***Abstract*—Messy Genetic Algorithm (GA) is applied to the satellite image clustering. Messy GA allows to maintain a long schema, due to the fact that schema can be expressed with a variable length of codes, so that more suitable cluster can be found in comparison to the existing Simple GA clustering. The results with simulation data show that the proposed Messy GA based clustering shows four times better cluster separability in comparison to the Simple GA while the results with Landsat TM data of Saga show almost 65% better clustering performance.**

Keywords—Genetic Algorithm: GA; Messy GA; Simple GA; clustering introduction

I. Introduction

As unsupervised image classification based on statistical methods, quantification theory and clustering are mentioned as typical methods [1]. Clustering methods can be broadly classified into hierarchical clustering, which treats pixels as individuals and classifies sets of individuals hierarchically, and non-hierarchical clustering, which divides a set of individuals at a time into a certain number of divisions [2]. The latter method is a method in which the initial cluster is given, the cluster to which the individual belongs is determined based on the distance between the cluster and the individual, the cluster centroid is obtained, and the individual is rearranged. The former differs from the latter in that clusters are formed based on the distance between individuals, between individuals and within and between clusters without giving an initial cluster [4]. In general, the latter method is frequently used because of faster convergence. In particular, when relocating, k clusters, i.e., k-means method [5] and ISODATA (Iterative Self Organizing Data Analysis Techniques A) [3] is famous.

In any clustering method, when n individuals are divided into k clusters, there is no guarantee that an optimal division result will be obtained. The latter can also be considered as a kind of optimal combination problem, and one of the effective methods is a Genetic Algorithm (GA) [6]. Clustering by GA, which effectively uses the effects of stochastic search and learning, is a method of improving the division by giving the evaluation criteria and initial division of the cluster. As a conventional method, there is clustering by Simple-GA [7].

However, since the location of the Simple-GA on the chromosome coincides with that on the chromosome, it is likely that the schema is superior or inferior depending on the location on the chromosome, and that the long schema is likely to be destroyed. On the other hand, Messy-GA is a variable-length list structure in which the chromosome-one gene expression is called a codon (locus allele). Therefore, it is unlikely that the schema will be superior or inferior due to the correspondence between the locus and the position on the chromosome, and the long schema can be expected to be preserved [7]. When applying the genetic algorithm to image clustering, cluster numbers are assigned by random numbers according to the pixel array, and the cluster number array (schema) effective for maximizing the fitness function is saved, crossed over, and mutation is performed. Probably searches for the optimal cluster while waking up, but originally the image has high spatial correlation, so the cluster to which the target pixel belongs is likely to match the cluster around it.

The author proposed a clustering method that takes such contextual information into account [8]. This paper further proposes a method using Messy-GA to guarantee schema preservation. The author takes up the degree of separation between clusters as the clustering accuracy, evaluate it using simulations and real satellite images (Landsat TM), and confirms the effectiveness of the proposed method.

II. Proposed Method

The author proposes clustering using Messy-GA in comparison with Simple-GA.

A. Messy GA Clustering

The schema length of the Simple GA is fixed. Therefore, relatively long schema which is effective for cross over is used to be broken. Consequently, it is difficult to find the most appropriate solution of chromosome. Meanwhile, cross over is much effective for Messy GA due to the fact that all the possible chromosome of maximum length can be prepared because the chromosome length is variable together with list of structural representation of chromosome. (1) Coding of chromosome, then initial pair of pixels number and cluster number is se.t (2) Fitness function evaluation. (3) Initialization. (4) Primordial phase. (5) Juxtaposition phase When the iteration number and the data number is exceed the threshold, all the pixels are assigned to cluster number.

B. Chromosome-Genotype Expression

Gene expression representing the state of dividing n individuals into k clusters is performed as shown in Table I.

Gene: C_i = 0,1,…, k-1

That is, a pixel is defined as an individual, and its cluster number is defined as a gene. Genes are arranged according to the order of pixel arrangement, and GA is used in an algorithm for stochastically searching for an optimal cluster of the pixel. At this time, the optimal cluster (remains) under the condition to maximize the fitness function shown as follows.

Gene sequence or a partial sequence of a certain part of the chromosome and another partial sequence (schema) at a certain probability, or by causing a mutation at a certain probability.

Find the best cluster. In the case of Simple-GA, since the schema description is fixed length, even if a valid schema for maximizing the fitness function can be searched, it is highly likely that it will be destroyed, but Messy-GA since the description of the schema is variable, the schema determined to be valid can be stored. This mechanism is shown in Fig. 1.

C. Fitness Function

The coding of the chromosome representing the division state in which n individuals are divided into k clusters is performed as follows:

$$((x_{i1}\ v_{i1})\ (x_{i2}\ v_{i2})\ldots\ (x_{in}\ v_{in})) \qquad (1)$$

where, assuming that the length of the chromosome in Simple-GA is 1,

$$1 >\ldots\ x_{i1},\ x_{i2},\ \bullet\ c,\ x_{in} >\ldots\ l \qquad (2)$$

and allow a variable length. x_{i1} ... x_{in} indicates a locus, and v_{i1} ... v_{in} indicates an allele value at the locus.

A fitness function is defined by equation (3).

$$S_B\ (k) = S_T\text{-}S_W\ (k) \qquad (3)$$

where S_T: sum of squares of n individuals,

$$S_r = \sum_{i=1}^{n} \|x_i - m\|^2 \qquad (4)$$

S_W (k): sum of square sums in a cluster of k clusters,

$$S_w(k) = \sum_{i=1}^{n} \sum_{x_i \in C_i} \quad \|x_i - m\|^2 \qquad (5)$$

S_B (k): sum of inter-cluster sum of squares of k clusters,

$$S_B(k) = \sum_{i=1}^{n} \sum \quad n_i\ \|m_i - m\|^2 \qquad (6)$$

The n individuals are divided into k non-empty, mutually exclusive clusters.

D. Selection / Selection Operation

The same number of chromosomes as the population of the previous generation are selected from the population of the previous generation by using the expectation strategy using uniform random numbers and elite preservation strategy according to the fitness. At the same time, selection is performed[1]. As an example, the decrease in the expected value in the expected value strategy is 0.75.

TABLE. I. GENE EXPRESSION

Pixel_No.	0	1	…	n-1
Gene(Cluster_No.)	C_0	C_1	…	C_{n-1}

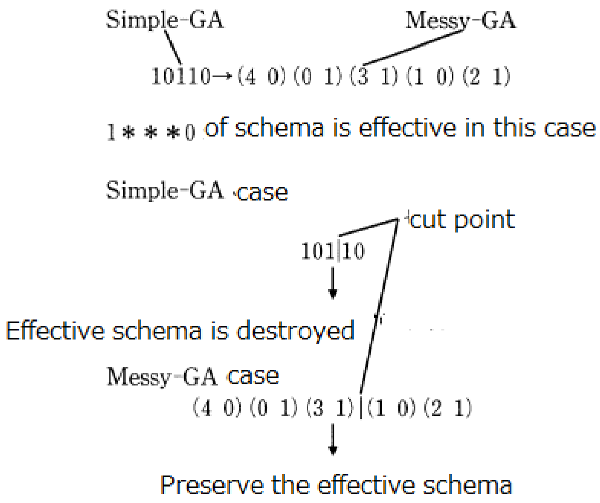


Fig. 1. Difference between Simple GA and Messy GA.

E. Crossover

The crossover operation performs multipoint crossover using dominant inheritance as a model. According to the crossover probability, cross-symmetric chromosomes are selected from the chromosomes selected in the selection and selection operation, and crossover is performed in the selected order.

At the time of crossover, the genotype mismatch between the two chromosomes occurs. Therefore, the reference locus is selected from the chromosomes selected as the crossover target using uniform random numbers. Based on the selected reference loci, the alleles are replaced according to equation (7), and the genotype matches.

$$j = i\text{-}h\ (\text{mod}\ n), \qquad i = 0,1,\bullet\ c,\ n\text{-}1 \qquad (7)$$

In this way, multipoint crossover is performed on chromosomes with unified allele types. However, actual allele replacement is performed only when the fitness of the replaced chromosome is improved. The chromosome where the allele replacement is performed updates the fitness every time the replacement is performed[2]. As an example, the crossover probability is 0.6.

F. Mutation Operation

According to the mutation probability, the chromosomal locus causing the mutation is randomly determined, and the allele is determined using the uniform random number at the determined chromosomal locus. If the fitness is improved when replacing with the determined allele, allele replacement is performed[3]. As an example, the mutation probability is 0.03.

G. Convergence Condition

The initial division of the cluster is set to 0 generation, and updating of the set generation is used as the program termination condition[4]. As an example, the end setting generation is 300,000 generations.

[1] As an example, the decrease in the expected value in the expected value strategy is 0.75.

[2] As an example, the crossover probability is 0.6.

[3] As an example, the mutation probability is 0.03.

[4] As an example, the end setting generation is 300,000 generations.

III. Experiment

The conventional method and the proposed method were applied to the simulation and actual satellite images, and the respective clustering accuracy was evaluated. Here, to show the superiority of the proposed method, the parameters of GA in the conventional method are the same as those of the proposed method.

A. *Simulation Parameters*

GA parameters are as follows.

- Early chromosome group 50
- Crossover probability 0.75
- Number of end generations 300,000
- Mutation probability 0.03 (Simple-GA only)

The simulation data creation parameters are as follows. Cluster individuals were generated by improving Neyman-Scott's method [9].

- 100 clusters・ 2 bands・ 2 clusters

Cluster average vector: $\mu_1 = (0.2, 0.9)^t$, $\mu_2 = (0.4, 0.6)^t$

- Cluster standard deviation $\sigma = 0.04$
- Distance between clusters 4σ

A population of 100 means an image of 10 × 10 pixels. 900 kinds of simulation image data were generated by changing the initial value of random numbers. Here, the distance between clusters i and j is shown in equation (8).

$$d_{i,j}^2 = (x_i - x_j)^t D_{i,j}^{-1}(x_i - x_j) + \frac{n}{2} ln|D_{ji}| \qquad (8)$$

where $D_{i,j}$ is the covariance matrix of i, j of the cluster, $|D_{i,j}|$ is its determinant, and n is the number of dimensions.

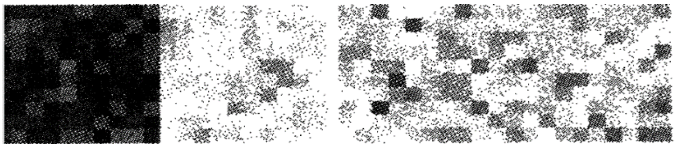
Normal random numbers were generated sequentially by giving the average and standard deviation, and constrained by the distance between clusters to construct a pixel array. Fig. 2 shows an example of the generated simulation image data. From the left, bands 1 and 2 of cluster 1 and bands 1 and 2 of cluster 2 are shown.

B. *Simulation Results*

Simulation images were generated to the extent that they could be considered as statistics (900 here). Clustering was performed by Simple-GA (SGA) and Messy-GA (MGA), and the degree of separation between clusters represented by equation (8) was evaluated. The cluster result images of Simple GA and Messy GA clustering are shown in Fig. 3, 4, respectively. Here the author shows only two of the 900 trials. At this time, Table II shows the number of generations up to convergence and the final degree of separation between clusters.

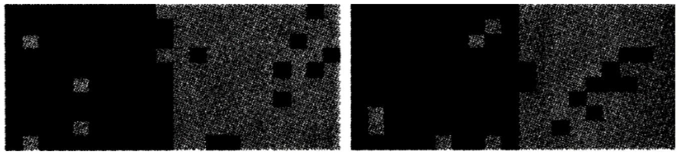
Fig. 5 shows an example of a change in the degree of separation between clusters (learning process).

Here, TBfitness and generation indicate the degree of separation between clusters and the number of convergent generations, respectively. In the figure, the broken line represents the learning process of SGA, and the solid line represents the learning process of MGA. All genotypes with deceptive order-length building blocks at the initialization stage because the chromosomes characteristic of MGA are of variable length and the chromosomes are composed of a list of loci and allele pairs can be generated and only genotypes with valid building blocks can be left to posterity. In addition, gene lists can be exchanged on the chromosome, and this optimization learning takes time. Chromosomes are significantly different from SGAs, which are represented by a fixed length.



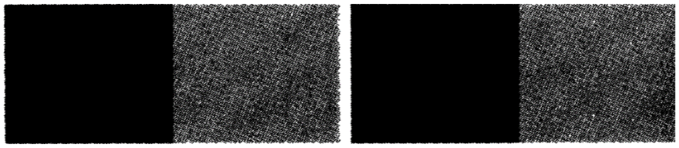
(a)Cluster#1Band 1 (b)Band 2 (c)Cluster#2Band 1 (d)Band 2

Fig. 2. Simulation Data used.



(a)Cluter#1 (b)Cluster#2

Fig. 3. Simple GA.



(a)Cluter#1 (b)Cluster#2

Fig. 4. Messy GA.

TABLE. II. The Number of Generations up to Convergence and the Final Degree of Separation between Clusters

Method	Separability_Between_Clusters	Iteration_No.
SGA	779.5	471
MGA	2866.1	6947

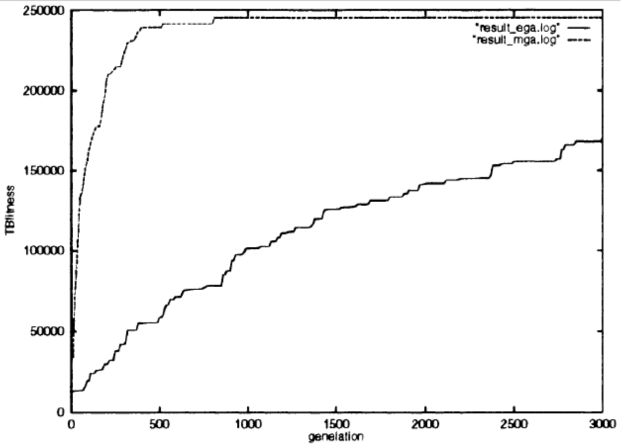


Fig. 5. An Example of a Change in the Degree of Separation between Clusters (Learning Process).

C. Satellite Image Results

Next, the results applied to real satellite images are shown. The GA operator parameters are as follows:

- Early chromosome group 50
- Crossover probability 0.75
- Number of end generations 300,000
- Mutation probability 0.03 (Simple-GA only)

The Landsat TM (Thematic Mapper) image around Saga city in March 25 1986 was used as an actual satellite image. Fig. 6 shows the location of intensive study area (Red circle) in the Google map.

From the image, a portion of 32 x 32 pixels in height and width is extracted as shown in Fig. 7. Landsat TM has the following seven spectral bands, including a thermal band:

Band 1 Visible (0.45 - 0.52 µm) 30 m

Band 2 Visible (0.52 - 0.60 µm) 30 m

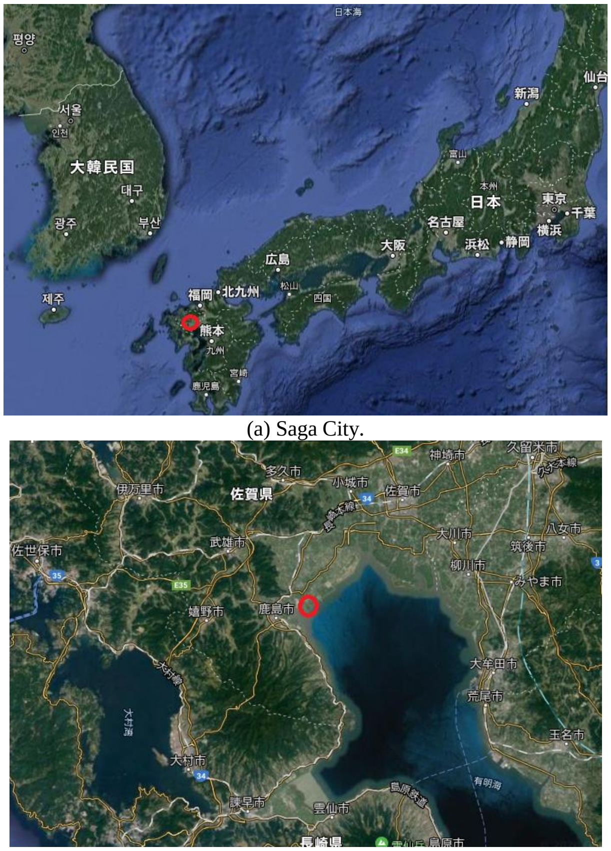
Band 3 Visible (0.63 - 0.69 µm) 30 m

Band 4 Near-Infrared (0.76 - 0.90 µm) 30 m

Band 5 Near-Infrared (1.55 - 1.75 µm) 30 m

Band 6 Thermal (10.40 - 12.50 µm) 120 m

Band 7 Mid-Infrared (2.08 - 2.35 µm) 30 m



(a) Saga City.

(b) Intensive Study Area.

Fig. 6. Location of Intensive Study Area.

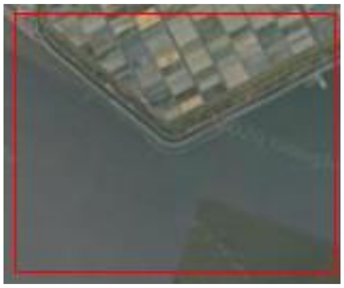


Fig. 7. Extracted Portion of Landsat TM Image.

Ground Sampling Interval (pixel size): 30 m reflective, 120 m thermal. 6 band data (excluding thermal band) are used for clustering. 5 clusters (urban, road, soi1, water, paddy) are assumed. Therefore, the number of clusters are set at five.

True color image of a portion of Landsat TM image which is acquired on March 25 1986 is shown in Fig. 8.

Fig. 9(a) shows only band 1 among the images used at that time. The resulting images by SGA and MGA clustering are shown in Fig. 9(b) and (c), respectively.

The clustered image of SGA shows noisy while that of MGA shows relatively smooth. In particular, nevertheless Ariake Sea area has to be clustered as one cluster, SGA result shows not only water body but also base soil, rice paddy, etc. Meanwhile, MGA result shows comparatively reasonable cluster.

Table III shows the number of convergent generations and the degree of separation between clusters.

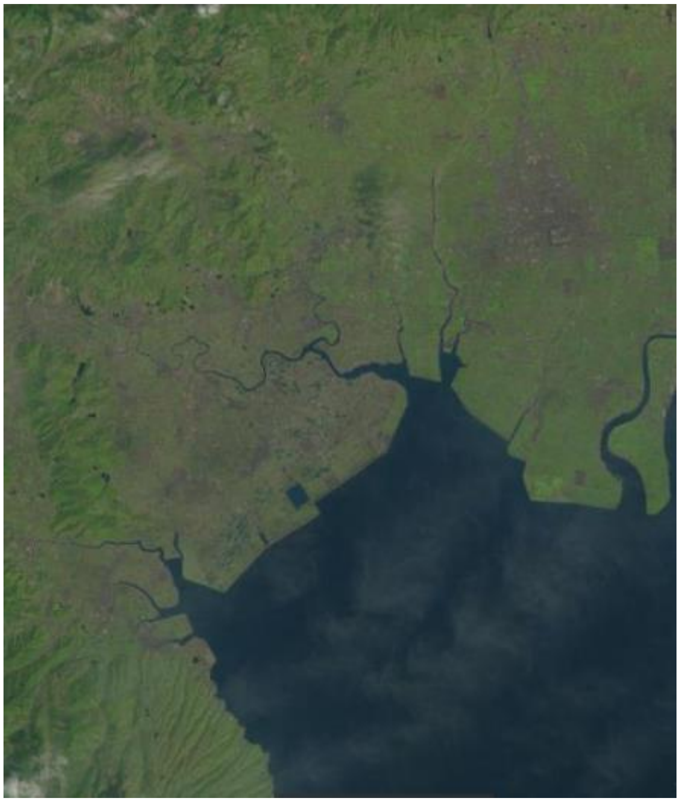
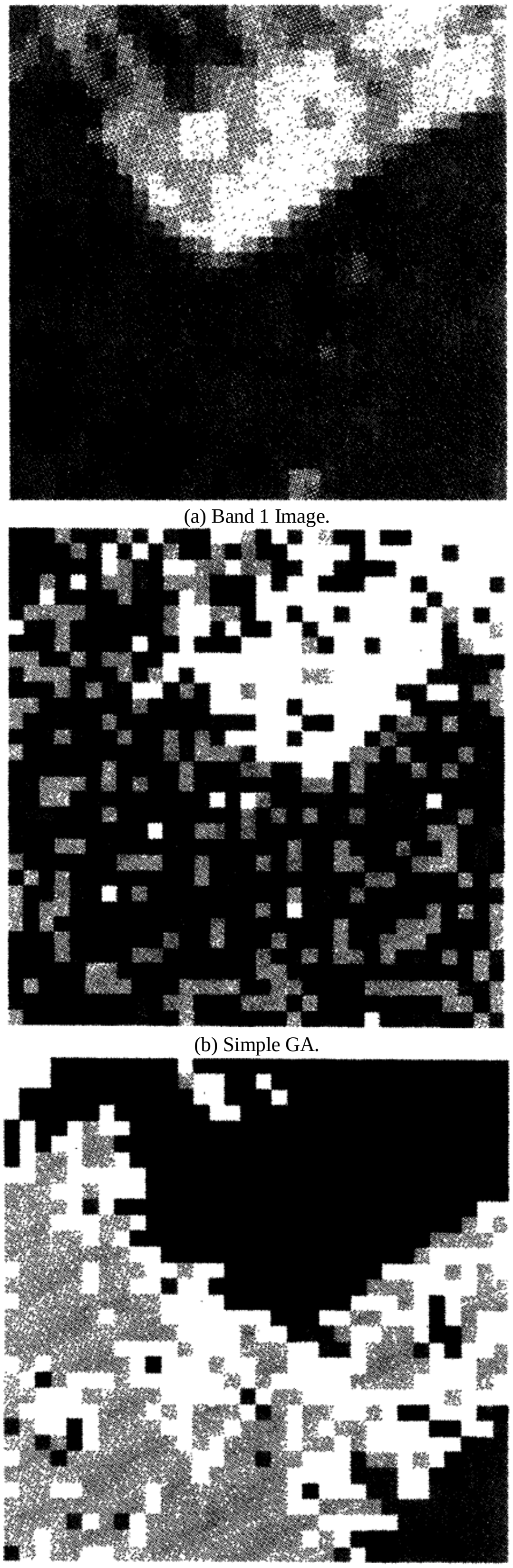


Fig. 8. A Portion of Landsat TM Image which is Acquired on March 25 1986.



(a) Band 1 Image.

(b) Simple GA.

(c) Messy GA.

Fig. 9. Band 1 of Landsat TM Image and the Resultant Images of Simple GA and Messy GA Clustering.

TABLE. III. CLUSTERING RESULTS USING LANDSAT TM IMAGES

Method	Separability_Between_Clusters	Iteration_No.
SGA	335	29238
MGA	554	299840

From these experimental results, it is found that Messy GA is superior to the conventional Simple GA from the view-points of reasonable clustered result and separability between clusters, the required time for clustering processes is much longer than Simple GA.

IV. CONCLUSION

Messy Genetic Algorithm (GA) is applied to the satellite image clustering. Messy GA allows to maintain a long schema, due to the fact that schema can be expressed with a variable length of codes, so that more suitable cluster can be found in comparison to the existing Simple GA clustering.

In Simple-GA, a gene has a fixed-length list structure, so a long schema is likely to be destroyed. In contrast, Messy-GA has a variable-length list structure and can store a long schema. As a result of comparing and evaluating the accuracy of the two clusters using 900 types of simulation data, the separation between clusters was shown to be about four times, and the result using Landsat TM image showed about 65% improvement, indicating that Messy-GA clustering turned out to be superior to Simple-GA. However, it was also found that the number of convergent generations was about 10 times higher for Messy-GA than for Simple-GA.

The author confirmed that both Simple-GA and Messy-GA surpassed the accuracy of k-means clustering, and also confirmed the tendency of accuracy improvement due to the difficulty of clustering, but the author will report these opportunities again.

V. FUTURE RESEARCH WORKS

Further experiments are required for validation of Messy-GA clustering effectiveness with the other remote sensing satellite images. Also, the applicability of the proposed Messy-GA clustering has to be attempted for not only remote sensing satellite image, but also the other images.

ACKNOWLEDGMENT

The author would like to thank former students of Dr. Akira Yoshizawa and Mr. Koichi Tateno as well as Professor Dr. Hiroshi Okumura of Saga University for his valuable comments and suggestions.

REFERENCES

[1] Mikio Takagi and Yuhisa Shimoda, edited by Kohei Arai, "Image Analysis Handbook", University of Tokyo Press, 1991.

[2] Anderberg, M.R., Nishida (translation): Cluster analysis and its application, Uchida Ritsuruho, 1988.

[3] Ball, G.H. and D.J.Hall: ISO-DATA-Novel method of data: Analysis and patter classification, Menlo Park, California, Stanford Research Institute, 1965.

[4] Rance, G.N. and W.T.Williams: A general theory of classification sorting strategies-Hierarchical System-, Cognitive Journal, 9, 4, 373-380, 1967.

[5] Selim, S.Z. and M.A.Ismail: K-mean type algorithms: A general convergence theorem and characterization of local optimality, IEEE Trans.on PAMI-6, 1, 81-87, 1984.

[6] T. Kato and K. Ozawa: Non-hierarchical clustering using genetic algorithm, IPSJ Journal, 37, 11, 1950-1959, 1996.

[7] Masashi Iba: Basics of Genetic Algorithms, Ohmsha, 1994.

[8] Satoshi Yoshizawa, Kohei Arai: Clustering using genetic algorithm combined with spectrum and context information, Journal of the Institute of Image Electronics Engineers of Japan, 31, 2, 202-209, 2002.

[9] J. Neyman and E.L.Scott: Statistical approach to problems of cosmology, Journal of the Royal Statistical Society Series B 20,1-43, 1958.

AUTHOR'S PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 37 books and published 570 journal papers. He received 30 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.htm

Improve Speed Real-Time Rendering in Mixed Reality HOLOLENS during Training

Rafeek Mamdouh[1*]
Ph.D. Candidate, Information Systems Dept
Faculty of Computer and Information Sciences
Mansoura University, Egypt
Assistant Lecturer, Multimedia & Graphics Dept
Faculty of Filmmaking and Performing Arts
BADR University
Cairo, Egypt

Hazem M. El-Bakry[2], Alaa Riad[3]
Information Systems Dept
Faculty of Computer and Information Science
Mansoura University, Egypt

Nashaat El-Khamisy[4]
Computers and Information Systems Dept
Sadat Academy, Cairo, Egypt

***Abstract*—Augmented reality (AR), virtual reality (VR), and mixed reality (MR) are advanced applications of computer visualization Which a hybrid structure allows users to explore novel used with other technologies in healthcare among other sectors give a promising future owing to the capabilities that come along with the technology that enables the medical personnel can carry out their surgical operations precisely. HOLOLENS 1, an MR product by Microsoft, is one of the first AR devices that have been widely applied in medicine for the treatment of complex diseases. It is also applied in operations that require a lot of care, for example, surgery of the liver. It is the main objective in the research to use HOLOLENS in performing surgeries while increasing Time Interval and controlling Semantic Segmentation while maintaining the truth of patient liver data during surgery for the segmentation liver 3d model. Next, we describe a new technology that increases the points of light, and the more the 3D intensity, the brighter the images the easier to interact with them. Holographic intensity is also to avoid blurring images to the point where the user sees through transparent HOLOLENS lenses. This improves *Time Interval* lens sensitivity and user detection in the environment. Finally, we describe a new framework for improving speed real-time render and model segmentation used hybrid Visualization between VR & AR called MR in which we decree render time speed through increase point light throw color calculations and energy function to be fast in sending and receiving data via WIFI unit.**

Keywords*—Mixed reality; time interval; semantic segmentation; Microsoft HOLOLENS; computer visualization

I. Introduction

Augmented reality (AR), virtual reality (VR) and mixed reality (MR) are the applications of computer visualization at different levels. There has been a mutation of technology starting from virtual reality via augmented reality and then to the recent mixed reality. Virtual reality immerses the user in a virtual world such that the user feels as if he or she is interacting with real-world objects [1]. This, however, is an illusion since the visualization is only made to behave like real objects. The applications of virtual reality can be seen in games like Pokémon go and training in some fields like engineering. Augmented reality took the second step after virtual reality by bringing in some reality, for example, the use of computed tomography (CT) scans that give a 3D transformation of an image as argued by [2]. This means that medical operations can be easily carried out due to the geometrical possibilities and ease of use. It has also extended mentoring and learning by giving efficient practical techniques to novice learners, for example, medics in the field of medicine and health.

MR is a conjunction between the AR and VR such that a real-world object can be maneuvered with the reference of a virtual object. In the medical sector, a patient can be operated in real-time using computer visualization the best being the HOLOLENS. Surgical operations that involve a lot of care, most notably, brain, liver, and other essential body organs can be carried out with a greater level of accuracy. The visualization from the eye of medical personnel is displayed on the screen which necessitates insights from other team members which will enhance accuracy further. Moreover, training of interns or novice personnel is made Easier since they can interact with augmented objects hence the most effective and fulfilled learning. This paper will focus on computer techniques and visualization regarding HOLOLENS and how it can be improved to enhance its Usefulness in liver surgery [3].

II. Background Research

HOLOLENS, specifically Microsoft HOLOLENS, is a mixed reality device that comes as a pair of smart glasses. Microsoft HOLOLENS is still under development and hence named Project Baraboo which was first released to the users in Canada and the United States. It runs under the operating system platform of windows 10 and is well known for being the first head-mounted device giving output as display and sound. The display is in the form of 3D images and hence easier to manipulate images as if they were real-world objects. Kinect was an inspiration that was used to come up with the tracking technology of the HOLOLENS which was the Xbox console, a gaming product, also by Microsoft.

Investors, that is, Asus in conjunction with Samsung, have shown interest in HOLOLENS and have pledged to improve the hardware part of the product alongside the concept that was

*Corresponding Author
https://orcid.org/0000-0002-9661-3551

used to come up with the product. Moreover, Microsoft has made the product accessible to the public through content management systems, for instance, GITHUB and Bit Bucket, to facilitate improvement from developers and programmers all over the world [1].

HOLOLENS was designed such that it can be tilted in an up and down or backward and forward manner by a cushioned headband. It requires to be mounted on the head and gives input in the form of a display which then can be adjusted by the headband which lies from the inside. The HOLOLENS has an adjustment wheel that a user can use to ensure HOLOLENS is comfortable in the head. The adjustment wheel distributes the weight that the whole unit has and also acts as support which acts as a prerequisite to the tilting toward the eyes at the front of the head. In comparison, unlike the backside that contains the adjustment wheel, the front unit consists of sensors, processors, projecting lenses and cameras among other related hardware. Also contained in the visor, which is tinted and consists of a pair of combiner lenses which is transparent acts as an output where images get displayed. The images are not displayed in the whole area of the visor but only the lower half. The calibration of the HOLOLENS is to the nearest IPD, which stands for inter pupillary distance, or it can be accustomed according to the user vision.

According to Joachimczak (2017), a pair of 3D speakers are located close to the ears of the user and distributed along the side edges. Audio speakers are competing with the typical sounds but give no obstruction to the sounds from the external environment hence the user will not hear the mix of virtual and environmental sounds, meaning there is no distraction at all. Binaural audio is generated by the HOLOLENS which is known to aid in the simulation of spatial effects. This means that the user can virtually perceive and estimate the location of a sound as if originating from a location that is from a virtual location or pinpoints [4].

The hardware part of the HOLOLENS consists of inertial measurement unit (IMU) [5] [6] Fig. 1, that is, the collection of a magnetometer, an accelerometer in conjunction with the gyroscope, four intelligent sensors distributed in pairs on every side, a depth camera which is energy-saving with a relatively wide angle of view which is 120 degrees by 120 degrees and photographic camera which can record videos. It also consists of an array of four microphones and light sensors with ambient characteristics.

HOLOLENS comes in handy with the HPU, which is a holographic processing unit that adds the functionality to the Intel Cherry SoC which contains the CPU alongside the GPU. HPU is a core processor that was solely manufactured or tailored to the HOLOLENS hence may not work in other hardware devices that do not have relative characteristics as the HOLOLENS. The HPU, just like the SoC, has LPDDR3 with a size of 1GB. Both HPU and SoC commonly share an SRAM of 8MB and in that case, SoC controls 64GB eMMC alongside operating under the windows 10 environment.

Moreover, the HPU utilizes 28 DSPs customized from Tensilica for processing and integrating the data received by the sensors alongside task handling techniques like recognition of gestures, spatial mapping, speech and voice recognition. A rechargeable battery is contained in the HOLOLENS with a lifespan of approximately 3 hours when actively used and approximately 2 weeks when left in standby mode. This shows that the HOLOLENS may not be used in an environment with no power supply, but the advantage is that once the power supply is adequate, the device can be charged while being used at the same time.

HOLOLENS supports IEEE802.11 ac which is Wi-Fi connectivity, one of the wireless local area networks (WLAN). It also features Bluetooth V4.1wireless connectivity which consumes less energy hence termed LE, low energy. Bluetooth LE is utilized in the headset for pairing with Clicker, which is an operating input device with the size of a thumb. The Clicker is used for scrolling the interface and selection of simulating objects when necessary. The clicker consists of a clickable surface for selection and sensors oriented which provides scrolling functions that may be in the form of panning or tilting the unit [3].

Elastic finger loop and USB micro-B version 2.0 are also contained in the Clicker which functions as device holder and internal battery charging respectively. Display and volume buttons for brightness are located at the top edge. Display buttons which are used to control brightness are near the left ear while volume buttons are on the right ear. There are adjacent buttons which are varying in shapes to make it easier for the user to distinguish when they touch, that is, concave and convex.

All these sums up to the HOLOLENS a Microsoft product that had several applications in the market. The HOLOLENS has its strengths and weaknesses which can be easily overcome with the advancement in technology and hence the reconstruction of the product to fit the required standards. The applications of HOLOLENS can be seen in windows 10 products like CORTANA, and other products like HOLOSTUDIO, HOLOTOUR, ACTIONGRAM, ROBORAID among other products. It has also been applied in various sectors of the economy, for instance, health, education, business and also in technological advancements.

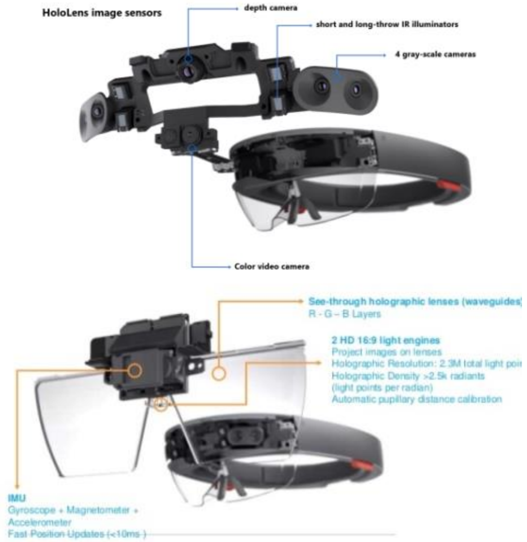

HoloLens image sensors
depth camera
short and long-throw IR illuminators
4 gray-scale cameras
Color video camera
See-through holographic lenses (waveguides)
R - G – B Layers
2 HD 16:9 light engines
Project images on lenses
Holographic Resolution: 2.3M total light poi
Holographic Density >2.5k radiants
(light points per radian)
Automatic pupillary distance calibration
IMU
Gyroscope + Magnetometer +
Accelerometer
Fast Position Updates (<10ms)

Fig. 1. Sensors on the HOLOLENS.

Mentoring or training novice medical students has been made easier with the HOLOLENS which gives CT scans that a student can learn at an instant. HOLOLENS, however, has some limitations, for instance, the lower storage, poor battery life, its limitation of usage to only the inside of a room or lab and poor rendering of images. This means that the HOLOLENS requires some adjustments to ensure that it performs as expected and to avoid disastrous effects that may come with the limitations it has. Microsoft team of developers issued the source code of the project to the public in GitHub to aid in the improvement of the project.

The use of Unity and reprogramming of the source code to ensure improved rendering and the readjustment of the device ergonomics will be a success to the project. The release of HOLOLENS is an improved version of HOLOLENS though it still has some limitations like rendering and speed or recognition of gestures in some instances.

III. Related Work

A. Augmented Reality (AR), Virtual Reality (VR) and Mixed Reality (MR)

Augmented Reality (AR) provides the user with the direct environment and combines additional data or information in the form of graphics amalgamated with the real objects. The object will then appear as an overlay and the user perception is somehow made to be an illusion by the device hence the user feels as if they are interacting with real-world objects [7]. Virtual Reality (VR), on the other hand, involves the simulation of a real or virtual environment which allows the user to interact unobtrusively that is as if the object is physically present. The only difference with AR is that VR does not alter the way the user views the object hence commonly referred to as a see-through technology as argued by Kowalski.

AR also has a wide variety of applications due to its easy interactivity and the way it brings the object to look as if they are real in conjunction with its cost-effectiveness. AR allows the creation of 3D shapes and can be manipulated virtually by applying geometrical operations like rotation, reflection, sheering among other operations. Evans claims that Mixed Reality (MR) is the technology that utilizes both AR and VR hence making the object to be easily manipulated, for instance, the HOLOLENS used in surgical operations and other engineering operations [8]. When all the three technologies are combined, they give groundbreaking solutions to the health sector in terms of efficient health services, education and the improvement of health operations altogether.

B. Applications of AR, VR, and MR in Medicine and Health Care

AR, VR, and MR have been applied in medicine and health care in several ways which include carrying out medical operations like surgery, medical records and scheduling, mentoring novice medical students and medical predictions among other several applications. Surgical operations are among the day to day activities of medical personnel that require maximum attention and accurate results. Chaotic environments like trauma sections are also difficult to be operated with the use of manual paraphernalia in the case where feedback is needed instantly [9].

The use of MR has facilitated the operations by allowing the professional to make decisions quickly and get insights to form the team from the visualization. HOLOLENS is an example of an MR device that has been widely applied in the health sector. Considering a chaotic environment, MR devices like HOLOLENS are suitable since they can be easily carried from one place to another without being connected to another device, that is, they are wireless [10].

C. Nature of Liver Surgery

The liver is the second largest body organ preceded by the skin. Liver surgery is one of the challenging surgical operations like those of the heart and the brain among other delicate body parts. This is because vital blood vessels that lead back and forth the heart have their way via the liver. Considering the liver itself, it is an organ that is fleshy hence can easily tear followed by serious bleeding when injured. Bloodless surgery is almost impossible with the liver though with the advancement in technology and the professional nature of medical personnel, bloodless bleeding can become almost probable [11].

According to Sharma, liver surgery, however, can only be carried out under some circumstances, for example, early stages of liver cancer or tumors in the liver. Liver transplant is also another instance where in some cases the whole liver is malfunctioned and hence the need to replace it altogether. The main advantage that comes with the liver is that it can regenerate and hence if the cancer cells have not spread enough in the organ, the affected part can be resection and the organ will continue to function normally [1]. Therefore, before liver re-sectioning can be performed, a lot of care should be taken and the high precision of the operation is demanded. This is where mixed reality technologies like HOLOLENS among other biotechnologies come into play.

D. The use of HOLOLENS in Liver Surgery

Due to the delicate nature of liver surgery, HOLOLENS comes in handy as a device that makes liver surgery easier and more precise [12]. Liver tumors or cancerous growths may involve a small area of the liver but require maximum care considering that the liver is highly vascularized and delicate to the point that a small surgical mistake could lead to serious repercussions. HOLOLENS is used to apply mixed reality technology such that a liver that needs operation is operated virtually and projected on a screen [13]. This allows other professionals to give insights at an instant while interacting with 3D objects that mimic the patient’s liver. The medical personnel may also be in different geographical locations but can give insights or even carry out operations themselves. With the use of the HOLOLENS, the bloodless surgery can almost be attained but considering the nature of the liver, the technology can only reduce the degree of blood in the surgery and also makes it convenient to apply surgical operations on patients.

E. Strengths of HOLOLENS

HOLOLENS is a mixed reality holographic computer by Microsoft and is far the best device that has a promising future for many fields including medicine [13]. HOLOLENS makes it easier for mentoring to be done by health professionals to novice medical students through the use of teleconferencing. It also allows medical professionals to get feedback from the team members and hence efficient medical operations. This reduces burnouts that usually occur in medics due to the maximum attention required to attend several patients. 3D images that are displayed in the form of CT scans may also be displayed on a screen for further checkups and attention [14].

F. Limitations of HOLOLENS

Although HOLOLENS has been embraced in various fields of economy and technology, it still has several limitations that if not taken care of may lead to serious damages to patients or losses in businesses. The hardware limitations are the major ones that involve heavyweight, batteries with a shorter lifespan, poor visual system in terms of rendering of images, low speeds in the use of gestures among other minute limitations [14]. Considering that HOLOLENS is connected to the internet, the medical data, for instance, may be exposed to malicious software like viruses and persons like hackers. In some instances, the HOLOLENS has to kill processes or apps to protect its memory and hence may lead to disastrous consequences taking the case of a patient undergoing operations that require maximum attention [15][16].

IV. Proposed Architectural Method

According to background research, HOLOLENS is an incredible holographic computer with great potential in technology. However, the device still needs some improvement in terms of architecture that comes in several forms, for example, hardware and software parts. The hardware may mainly deal with optimizing the components that will at the same time be compatible with the software. The software is to ensure that the functionality of the hardware is presented to the user in a platform efficient and easy to interact with. The following are the proposed architectural designs that should be made for HOLOLENS to be optimized and hence achieve its maximum potential. The main focus is the rendering techniques, ergonomics, and the best user experience as much as possible.

Increasing the light points for resolution boost in conjunction with the number of light concentrated per radian area for density improvement Fig. 2. Currently, HOLOLENS considers holographic density more than the light points which make the images brighter but the images seem blurred. Light points are created from a combination of colors mainly red, green and blue, the RGB colors. The more the light points, the higher the holographic density hence brighter images that are easy to interact with. Considering that holographic images come in 3D form, it means brightness is essential for good user experience. HOLOLENS was created to optimize a holographic density approximated to 2.5K radians and the light points to be applied were 2.3 million. Increasing the number of light points will increase the efficiency of HOLOLENS especially attacking robots which takes less coverage of the screen where HOLOLENS has more advanced functions such as gaze, spatial, sound and mapping design working by MRTK [17]. Holographic intensity from the HOLOLENS is also to avoid images getting blurred to the point where a user sees through the transparent lenses of HOLOLENS. This improves the sensitivity of the lens and detection of the user in the environment. The conversion of 2D to 3D together with the detection of voice and data is recorded as shown in the flowchart below Fig. 3 and Fig. 4.

Lenses are what magnify the images to be interacted with while the field of view determines the range or area that the eye can see or interact with images. Increasing the field of view of HOLOLENS lenses while at the same time maintain the quality of images displayed will be an optimizing step to visualization. This means more hardware maneuvers and software reconstruction to go in line with the specification. Most users who have used HOLOLENS described the field of view as rectangular which means that what can be observed depends on the rectangular area alone. Adjusting the vertical view of the lenses will make increase the field of view giving the user a clearer vision as they interact with holographic images. The ability to immerse, where the user gets deep into the holographic world, is what the users demand considering the use in situations such as surgery or observation of tiny cancerous growths that require clearer vision alongside coupled with maximum care.

The field of view reported being rectangular which makes the user see fewer HOLOLENS can be improved by using remote middleware of the holographic HOLOLENS connected to the PC by the Unity engine which displays images in 3D. The PC with high performance takes most of the workload and renders the images, passes them to Wi-Fi connection then which is then received by the HOLOLENS in frames form. Since the workload has been reduced it just displays, captures, and transmits information interactively which includes gestures alongside voice. The Geographical Information System (GIS) [18] collects information from the environment in real-time. The holographic remote player adjusts the field of view accordingly depending on the scenes either large or small. The pictures below give a demonstration of this concept in Fig. 5.

This is the behavior during smaller scenes. It can be noted that during the small scenes there is no need for a holographic remote player since the applications can be deployed directly to the HOLOLENS devices Fig. 6.

Adjusting the HOLOLENS to allow greater pupillary distance so that the user does not get eye strains or motion sickness which may come when the holographic computer is used for a long time. Users have complained that the pupillary distance is too short especially considering some users are long-sighted while others are short-sighted. When the HOLOLENS is moved to a considerable distance, the images get blurred and rendering problems set in. In some cases, if a user moves the HOLOLENS too close to the eye, the holographic images disappeared meaning the user is inconvenienced. Adjusting the pupillary distance to a considerable distance that can fit the nature of the user's eye will create a good experience for the user. Considering biometric techniques to the eye such that HOLOLENS will adjust according to the movement of the user's pupil and in conjunction with HOLOLENS lenses will ensure accurate and customized images to the user.

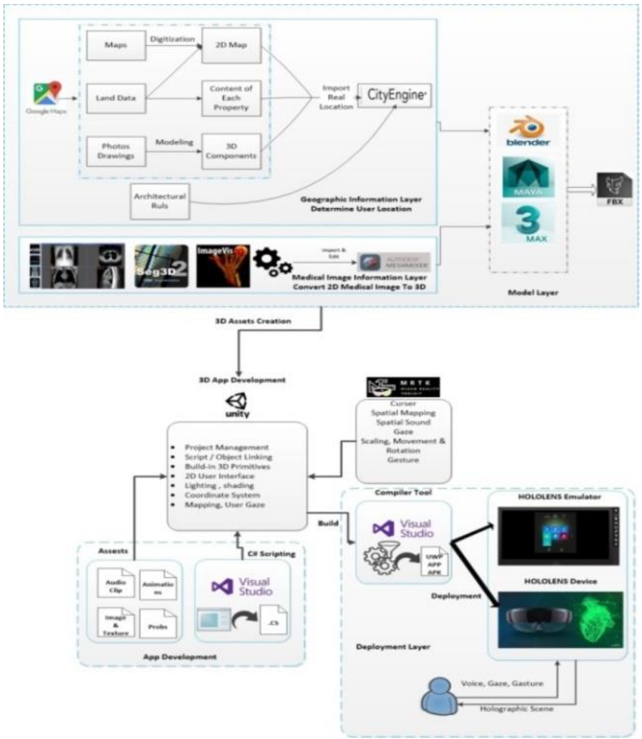


Fig. 2. Proposed Architectural: Increasing the Light Points for Resolution.

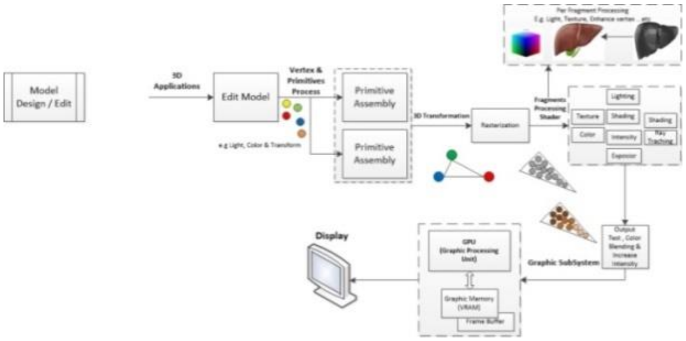


Fig. 3. The Algorithm Flowchart of Convert 2D Medical Image To 3D Model.

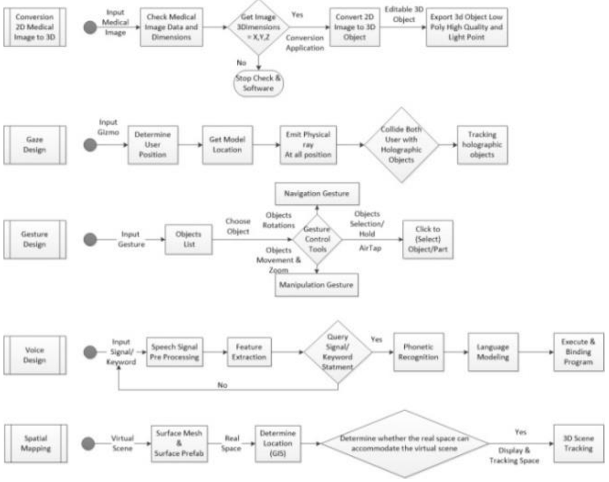


Fig. 4. The Conversion of 2D to 3D together with the Detection of Voice and Data is Recorded.

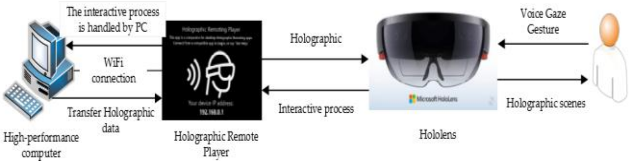

The interactive process is handled by PC
WiFi connection
Transfer Holographic data
High-performance computer
Holographic Remote Player
Holographic
Interactive process
Hololens
Voice Gaze Gesture
Holographic scenes

Fig. 5. HOLO Application for (Large Scene) throw WIFI Connection.

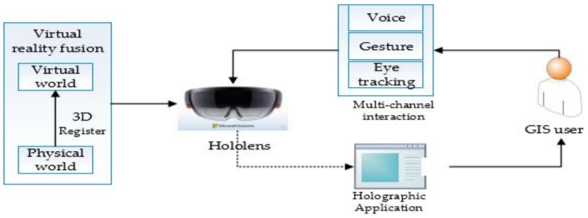


Fig. 6. Holo Application for (Small Scene).

Increasing the number of GPUs alongside the CPU and HPUs so that images are rendered much faster and with a lower workload to the motherboard of the HOLOLENS. The HOLOLENS has a motherboard that contains all the hardware meaning the space is limited. Unlike Oculus Rift, HOLOLENS is a computer that does not require to be mounted, that is, a standalone computer. This means the rendering of games may be quite difficult due to limited storage making some game lovers lose interest when their games fail to load as expected [10]. Increasing the number of GPUs with the HPU will ensure the rendering of quality images and videos. With the minimum motherboard space, it will require the minimization of the size of the hardware and at the same time maximizing the storage space. Fig. 7 below diagram shows how the GPU and CPU perform the tasks in HOLOLENS.

Increasing the number of FPS to the HOLOLENS will also optimize the images and data rendered to the screen. The frames per second of the HOLOLENS at the moment are 60 fps meaning when a user takes a preview of the image, it drops to around 40 fps hence the image appears blurred. If the GPU is increased together with the HPU tailored for the HOLOLENS, the fps will be possible with the increased processing power. Therefore, what should always be considered is the maximum number of frames per second that can be comfortably supported by the motherboard of the HOLOLENS. 80 fps can be comfortably supported with the increased GPU and HPU also with the more space created if the data is streamed to the cloud using Wi-Fi connectivity.

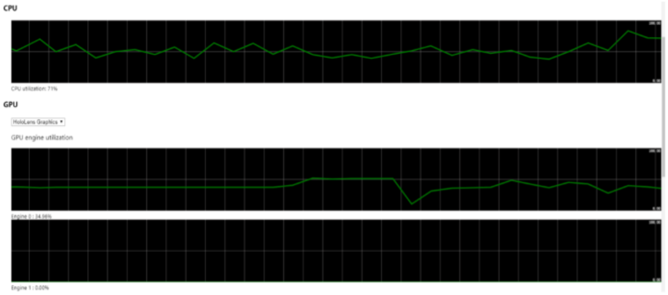


Fig. 7. CPU and GPU usage Percentages: Deployed on PC HOLOLENS.

Streaming images, video and data to the cloud using the IEEE 802.11ac Wi-Fi to create more space for user interactivity and image rendering [19]. All the other limitations like rendering, increasing the number of light points among others but with the limited space, the applications will often crash, to preserve the HOLOLENS mostly reported in the way HOLOLENS kills some activities when it reaches the extent of low storage to preserve itself. The limitations of storage discourage HOLOLENS usages in situations like hospital theaters where precision is compulsory and much attention demanded [20]. The strength that the HOLOLENS comes is its ability to connect to the Wi-Fi network meaning the data can be streamed to Microsoft Azure, the cloud platform by Microsoft. It will make HOLOLENS a device to be embraced by everyone since more storage space for CPU to perform tasks will mean quality holographic images, a large number of applications to be supported among other features. It may, however, be limited by cloud storage but Microsoft will have to make cloud storage affordable so that a large number of users can afford to enjoy the capabilities of improved HOLOLENS.

V. Results

The resulting proposed HOLOLENS will be a computer that is agronomical such that the user finds it comfortable when using it. Creating more storage space by allocating space in the cloud to the data in the HOLOLENS will make it a reliable computer to be used in a wide variety of places without the fear of disastrous effects. Visualization in the HOLOLENS is more concerned with the brightness of the holographic images but with the proposed system, the clarity of the images is also put into account so that a user can easily immerse themselves in the composite world. Wi-Fi storage came with the HOLOLENS since the beginning but the strength of the Wi-Fi, when combined with the cloud storage in streaming the images, will tap its greatest potential. A final product with a clear vision, improved field of focus, adjustment of pupillary distance and improved image rendering will make HOLOLENS a device to be embraced by companies and investors Fig. 8.

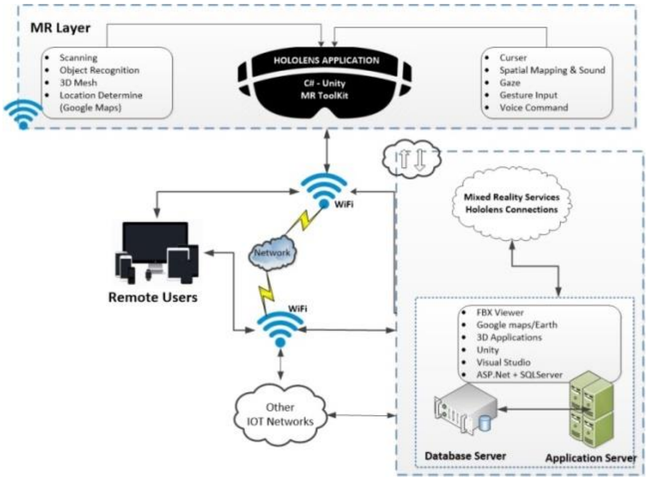


Fig. 8. Wi-Fi Storage with the HOLOLENS Combined with the Cloud Storage.

VI. Conclusion and Future Work

Indeed, HOLOLENS is a device that leads the computer technology to another era of augmented age. It is a fantastic device that has already been applied in some sectors though it is still at the testing stage. However, the HOLOLENS has many limitations which can be improved although some improvements may involve a lot of expenses for the company. Nonetheless, if the device is improved as proposed, the users will increase who will then increase the profit margin. The holographic computer has higher demand in institutions like health and education owing to its ability to immerse the user in a virtual world where they can retain a lot of information when compared to reading or watching a video. In the hospital operation room, especially for treatments like cancer which involves microscopic organisms, the HOLOLENS can combine the virtual reality with the real world giving the medical personnel the needed feedback from colleagues and efficient treatment of patients.

Abbreviations

- 3D – Three Dimensions. The type of dimensional geometry involves length, width, and height, unlike 2D which consists only of length and width.
- AR – Augmented reality. The technology that advances VR in that the user not only intersects with a limited number of objects but a wide variety including geometrical operations like rotation among others. AR superimposes images created by the computer to the real world making the user have a composite view.
- C# - C sharp - A programming language that has the same characteristics as python and other object-oriented programming languages.
- CT – computed tomography. This is a tomography that comes in the form of a computer controlling the motion of scans and x-ray detectors along with the source. CT scan takes images, processes them and give desired output in the form of a holographic image.
- eMMC –Embedded Multimedia Card - A technology that acts as storage for most mobile phones and related devices. On the phone, it is the internal storage that comes with the device apart from the external storage that someone can add up the storage.
- Fps – frames per second. The number of varying images or data that can be handled at once by a camera.
- GB- Gigabytes. The storage measurement capacity equivalent to approximately 1000MB.
- GPU- Graphics Processing Unit. The computer processor dedicated to rendering graphic images for quality graphics to be displayed.
- HPU- holographic processing unit. One of the dedicated processors tailored for processing images in holograms especially HOLOLENS.

- IEEE802.11 – institute of electrical and electronics engineers. It is the framework for the Wi-Fi network.
- LE – low energy.
- LAN – Local area network. A network that gives connectivity to only a certain radius in an area, unlike the internet that covers the whole world.
- LPDD3 –Low Power Double Data Rate.
- MB- Megabytes. The memory measurement capacity equivalent to 1000 KB.
- MR – Mixed reality. Combination of AR with real-world objects making objects look as though they are real.
- OOP – Object-Oriented Programming. The type of programming that employs the use of objects instead of structures. OOP separates the object from the model and view making it easier to upgrade anytime without a lot of readjustments.
- RGB-Red, Green, Blue
- SRAM – Static Random Access Memory.
- VR- Virtual reality. The technology in which a user is immersed in a virtual world and interacts with virtual objects.
- Wi-Fi – Wireless Fidelity. One of the LAN networks that conform to the IEEE802.11 ac framework
- MRTK - Mixed Reality ToolKit
- IMU - Inertial Measurement Unit

Compliance with Ethical Standards:

1) Funding: This study not funding

2) Conflict of Interest: No conflict exists: The authors declare that they have no conflict interests in this work.

REFERENCES

[1] Chen, Henry, et al. "3D collaboration method over HoloLens™ and Skype™ endpoints." Proceedings of the 3rd International Workshop on Immersive Media Experiences. ACM, 2015.

[2] Catanzarite, Joshua B., and Ryan J. Schoenefeld. "Patient-specific computed tomography guides." U.S. Patent No. 9,066,727. 30 Jun. 2015.

[3] Sauer, Igor M., et al. "Mixed Reality in visceral surgery: development of a suitable workflow and evaluation of intraoperative use-cases." Annals of Surgery 266.5 (2017): 706-712.

[4] Joachimczak, Michal, Juan Liu, and Hiroshi Ando. "Real-time mixed-reality telepresence via 3D reconstruction with HoloLens and commodity depth sensors." Proceedings of the 19th ACM International Conference on Multimodal Interaction. ACM, 2017.

[5] Dawid Borycki," Programming for Mixed Reality with Windows 10, Unity, Vuforia, and UrhoSharp "Chapter 1, Introduction to Windows Mixed Reality, 1st Edition, Microsoft corporation by Pearson Education Int. (2019)

[6] Clemente Giorio, R&D Senior Software Engineer,"HoloLens and Windows Mixed Reality",Feb 20, 2017,"https://www.slideshare.net/tinux/hololens-and-windows-mixed-reality":P.18:25 (accessed on 21March 2020)

[7] Kowalski, Marek, et al. "Holoface: Augmenting human-to-human interactions on hololens." 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018.

[8] Evans, Gabriel, et al. "Evaluating the Microsoft HoloLens through an augmented reality assembly application." Degraded Environments: Sensing, Processing, and Display 2017. Vol. 10197. International Society for Optics and Photonics, 2017.

[9] Swayze, Jeffrey S., et al. "Surgical system with augmented reality display." U.S. Patent Application No. 15/383,004. 2017

[10] Colenbrander, Roelof Roderick, et al. "Dynamic network storage for cloud console server." U.S. Patent Application No. 14/848,197. 2017.

[11] Sharma, C., et al. "Visual gaze patterns reveal surgeons' ability to identify the risk of bile duct injury during laparoscopic cholecystectomy." HPB 20 (2018): S239.

[12] Adabi, Kian, et al. "Optimizing measurements in plastic surgery through holograms with Microsoft Hololens." Plastic and Reconstructive Surgery Global Open 5.9 Suppl (2017).

[13] Coppens, Adrien. "Merging real and virtual worlds: An analysis of the state of the art and practical evaluation of Microsoft Hololens." arXiv preprint arXiv:1706.08096 (2017).

[14] Coffey, Dane, et al. "Computer visualization of anatomical items." U.S. Patent No. 9,818,231. 14 Nov. 2017.

[15] Kress, Bernard C., and William J. Cummings. "11-1: Invited paper: Towards the ultimate mixed reality experience: HoloLens display architecture choices." SID symposium digest of technical papers. Vol. 48. No. 1. 2017.

[16] Kress, Bernard C., and William J. Cummings. "Optical architecture of HoloLens mixed reality headset." Digital Optical Technologies 2017. Vol. 10335. International Society for Optics and Photonics, 2017.

[17] Mixed Reality Toolkit-MRTK. Available online: https://microsoft.github.io/MixedRealityToolkit-Unity/Documentation/Input/Gaze.html (accessed on 21 March 2020)

[18] Wang, Wei & Wu, Xingxing & Chen, Guanchen & Chen, Zeqiang. "Holo3DGIS: Leveraging Microsoft HoloLens in 3D Geographic Information. ISPRS International Journal of Geo-Information." 7. 60. 10.3390/ijgi7020060., (2018).

[19] Hu, Feng, Siqi Liu, and Libiao Jin. "The algorithm of channel estimation based on IEEE802. 11ac." 2017 3rd IEEE International Conference on Control Science and Systems Engineering (ICCSSE). IEEE, 2017.

[20] Müller, Christoph, et al. "Interactive molecular graphics for augmented reality using HoloLens." Journal of integrative bioinformatics 15.2 (2018).

Model of Tools for Requirements Elicitation Process for Children's Learning Applications

Mira Kania Sabariah[1]
Department of Electrical and Information Engineering
Universitas Gadjah Mada, Yogyakarta
Indonesia
School of Computing
Telkom University, Bandung
Indonesia

Paulus Insap Santosa[2]
Department of Electrical and Information Engineering
Universitas Gadjah Mada, Yogyakarta, Indonesia

Ridi Ferdiana[3]
Department of Electrical and Information Engineering
Universitas Gadjah Mada, Yogyakarta, Indonesia

***Abstract*—Requirements Elicitation are the initial stages in the application development process, where a set of needs from the system will be built and obtained by communicating with stakeholders who have a direct and indirect influence on those needs. Failure in the requirements elicitation process was caused by weak communication. Communication is an essential thing in carrying out the requirements elicitation process. The selection of the right elicitation technique is not only a solution. Informants as sources of information on requirements also need to be considered. The choice of the correct technique often fails because of the tools not useful. The availability of the right form of equipment needs to be considered so that the communication between the elicitation team and the informant goes well. Children have characteristics not the same as adults. Limitations in terms of psychomotor, cognitive, and emotional children are considered in choosing elicitation techniques and tools. These limitations are also influenced by the age range of child development. The use of digital elicitation devices is recommended to be used in the requirements elicitation process. The presentation of interactive tools makes it easier for children to convey their desires. In learning applications for children, aspects of pedagogy that need to be explored are learning styles and children's thinking abilities. Every child in every age range has a different preference for learning style. That is because children do not have learning experiences. That also applies to the level of thinking ability of children. Therefore, these two things need to be appropriately explored when the learning application development process. The proposed elicitation tool model was made by taking into account both components of that pedagogical aspects. The test results of the built model show that the application has satisfaction. That means that children can communicate well in conveying the needed as requirements to the learning application.**

***Keywords*—Requirements elicitation; communication; children learning application; pedagogical aspect; learning style**

I. Introduction

Requirements elicitation are the initial stages in the application development process, where a set of needs from the system will be built and obtained by communicating with stakeholders who have a direct and indirect influence on these needs [1]. According to Rupp, 60% of failures in the software development process occur due to requirements elicitation [2]. Failure in the requirements elicitation process was caused by difficulties in communicating between humans[3]. Ambiguity in communication often causes obstacles in transferring knowledge that causes documentation of needs to be incomplete and clear [4]. Communication is a relational process in creating and interpreting a message to get a response [5]. Communication is not just an expression but also persuasion, control, and influence in meetings between two people, or communication between two people with the existence of feedback and the role of the speaker and listener alternately. Then an interaction occurs [6].

Documentation of needs tends to be in the form of face to face communication [7]. Other problems, if there is no documentation related to software requirements specifications, will cause the requirements quality assurance (QA) process of the application to be built into a difficult one [8]. In many cases, the selection of elicitation methods or techniques was not based on application content or the strength of elicitation techniques but only based on a tradition, or that is deemed familiar by the developer [9].

The requirements elicitation process can run effectively if the developers have good expertise and knowledge in choosing an elicitation technique [10]. In many cases, the selection of elicitation methods or techniques was not based on application content or the strength of elicitation techniques. Sometimes only based on a tradition or that is deemed familiar by the developer [9]. Also, many developers in practice do not pay attention to requirements elicitation techniques [11]. Another thing that causes the failure of a requirements elicitation technique is the way of communication that is not by the level of knowledge of users [12], [13]. User involvement in the elicitation process is a factor of success in selecting elicitation techniques [14] and fulfills the usability of the applications built [15].

In the construction of children's learning applications, children's involvement in the elicitation process was almost never done. The current phenomenon, the involvement of children as users in the application development process, is often emphasized only at the testing stage so that sometimes the idea of the application being built is still determined by the developer [16]. By involving children in the elicitation process, of course, we need to pay attention not only to their elicitation techniques but also to consider the tools used in exploring the needs of the applications to build. Also, the direct involvement of children in the elicitation process can have a positive impact on children in using the learning products produced [17].

The objectives of this research are two, including (i) identify the form of media the right communication tool to assist the elicitation process, (ii) design a model of communication tools that will be used in the elicitation process according to the recommended media form at the first destination.

II. Literature Review

User requirements are the highest level in requirements [18], and their contents consist of a set of user desires [19]. Failure to define user requirements will impact the quality and satisfaction of the user of the application. Another thing also has an impact on not achieving the objectives of the software. Elicitation requirements are an initial step in defining requirements, one of which is user requirements. Communication is a general cause of failure in the requirements elicitation process [20].

Communication is a relational process in creating and interpreting a message to get a response [5]. In conducting communication, it is necessary to have feedback and the roles of the speaker and listener who alternately [6], which then occurs an interaction. The interaction process can occur if the message conveyed is understood by both parties. Piaget's convey [17] that children have limited ability in communication in every age range. Cognitive, psychomotor, and emotional are the factors that influence limited ability [18]. These factors certainly affect the requirements elicitation activities.

The different cognitive, psychomotor and, emotional development is certainly a consideration in choosing elicitation techniques. Interview and prototype techniques [21], are recommended as appropriate techniques for child respondents. The selection of appropriate techniques needs to be followed by appropriate tools so that communication can go well. The development of technology has a positive impact on children. This is evidenced by the many uses of technology in children, especially in the learning process [22]. Technology is considered as an interactive media for children in the learning process. Technology that can present audio and visual forms is considered quite effective and efficient in helping children interact.

Effective and efficient interactions and issuing of valid results are needed in the requirements elicitation activities. Agile methodologies that are widely used in the software development process today require these conditions [23]. Rapid iteration and relatively short development time [24], of course, requires the application of an appropriate technique and tool. Documentation of requirements is also a demand that is needed so that the verification and validation process can be done quickly and correctly. The application of digital technology as a tool in conducting requirements elicitation is widely used today. This concept is widely applied to the requirements elicitation framework [19], [25].

III. Methodology

A. Methods

The research was conducted in two stages. The reason for the two stages is because there are two objectives to be achieved. In the first stage, the data collection process was carried out using interview techniques and literature studies. Interview techniques were carried out using questionnaires and interviews with child learning experts. The questionnaire used was paper and digital. Interviews with experts were conducted to determine the content presented in the questionnaire. In addition to interview techniques, literature studies were also carried out. The keyword was used in the literature review are learning applications, children's characteristics, elicitation requirements, and elicitation techniques. The use of interview techniques and prototypes was based on the results of previous studies related to the best elicitation techniques used when communicating with children [20]. In the second stage, interviews and prototype techniques were carried out. The prototype was built according to the model produced from the results of stage one.

B. Participant

Participants who will be involved to achieve the first goal are children aged 6-8 years with primary school level education 1-3. To achieve the first goal, participants consisted of 33 children with a gender composition of 18 girls and 15 boys from 3 elementary schools. Meanwhile, to achieve the second goal, 32 participants were involved. The number of participants is adjusted to the limit of quantitative research [21], as many as 30 children.

C. Material

The material used in this application is a questionnaire and prototype applications. Two type of questionnaires that will be used in this research. The first questionnaire was given when determining the right form of tool to communicate with children. The second questionnaire was created to measure user satisfaction from the proposed model of tools.

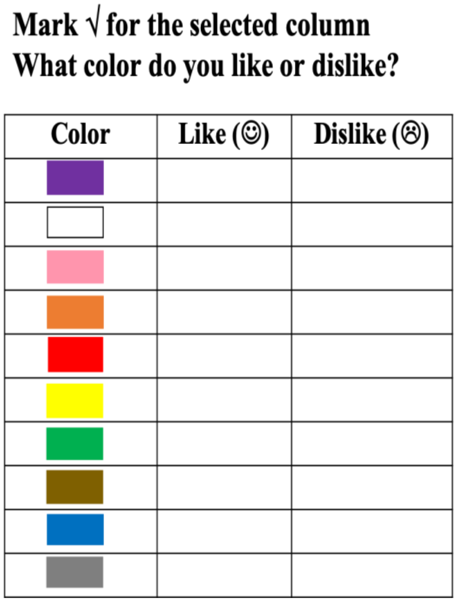
The questionnaire was made to answer the first purpose of this study, which is related to the form of tools that are appropriate for the child's respondent. There are two forms of questionnaires to be used in the data collection process. First is a paper questionnaire, and the second is the digital questionnaire. The questionnaire in digital form was presented in the form of an application.

The content of the two questionnaires is the same, where the difference is only in the form of presentation. The content presented is adjusted to the age of the development stage of the child 6-8 years, which refers to Piaget's. Questions were presented in four types of content, namely color, color in geometry, 2D / 3D geometry, and pictorial objects. Fig. 1 is a sample questionnaire in the form of a paper, and Fig. 2 is one example of a questionnaire in digital form (application). The children were asked to rate each question in the questionnaire with two types of answer choices, namely likes and dislikes. Questionnaires in digital form were built in the form of mobile-based applications. The second questionnaire was made to measure satisfaction from the proposed tool. The questionnaire presents nine questions (P-1-P9) related to user satisfaction.

Among them are the ease of using the application, ease of using each navigation, helps in choosing colors and objects, ease of reading texts. Likert scale (1-5) was used in the questionnaire.

D. Measure

Data measurement results from the first questionnaire were conducted using a non-parametric test with the Wilcoxon approach. This approach was carried out to see the difference between the selection of paper and digital form based on the time value. The second questionnaire using the analysis interval to processing data results.



Mark √ for the selected column
What color do you like or dislike?

Color	Like (☺)	Dislike (☹)
🟪		
⬜		
🟥		
🟧		
🟥		
🟨		
🟩		
🟫		
🟦		
◼		

Fig. 1. Example of a Paper Questionnaire.

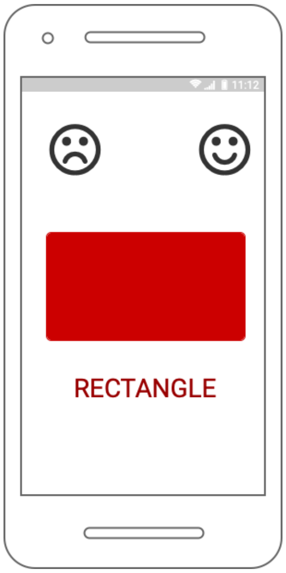


Fig. 2. Example of a Digital Questionnaire (Application).

IV. Result and Discussion

This research is divided into two stages, where the first stage is identifying the form of elicitation tools and the second stage is modeling the tools.

A. Identification the form of Requirements Elicitation Tools

Data retrieval related to the determination of the form of requirements elicitation tools for children's learning applications was carried out on 33 children. Children fill out questionnaires that were presented in the paper and digital forms (apps). Table I shows the results of the processing of the two forms of questionnaires measured based on the interaction time of each respondent in answering each question. Each material was presented in the form of paper and digital. From Table I, there are several things that can be concluded that communication with children tends to be more productive using applications (digital) compared to paper.

The difference between the use of paper questionnaires and digital has a time difference of about 5.91 seconds with an answer correlation value of 0.71. That was also proven by conducting a non-parametric test using the Wilcoxon approach with 0.003, as shown in Table III. The results can be concluded that there is a difference between the paper and digital approaches. Thus, the digital approach is more recommended in the process of needs elicitation with child respondents because it has a faster time, as shown in Table II. The use of digital media is also considered to improve children's understanding of the learning process, especially if it was presented in cross-platform forms [22].

B. Proposed Model of Tools for Requirements Elicitation

Based on the results of data retrieval through interviews and prototypes of the form of instruments recommended in the needs elicitation process is tools in the form of mobile-based applications. The model of elicitation tools built must be adapted to the requirements that need to be explored in building children's learning applications. There are two types of applications that tend to be made for children's education applications based on the results of interviews with five child education application developers in Indonesia. The type of application is in the form of games and simulation (non-game). Both types of applications have different characteristics and approaches to the development process so that the impact on the elicitation application model will be built. Content or problem domain becomes the primary key in the selection to determine the type of application to be made according to the requirements modeling language (RML) approach [23]. The other most crucial component to consider in the learning process is learning styles. VARK learning styles tend to be recommended in children’s learning applications. This is because VARK learning styles define learning strategies according to children’s sensory preferences, namely, visual, auditory, reading/writing, and kinesthetic [24]. In addition, children in the learning process do not have learning experiences. The content that will be presented in the form of VARK learning styles is also differentiated based on the thinking skills level that refers to Bloom’s theory [25].

TABLE. I. THE RESULTS OF PROCESSING QUESTIONNAIRE DATA USING PAPER AND DIGITAL

Respondent	**Paper**	**Digital**	**Paper**	**Digital**	**Paper**	**Digital**	**Paper**	**Digital**
	Color	**Color**	**Color Geometri**	**Color Geometri**	**Geometri 2D/3D**	**Geometri 2D/3D**	**Pictorial Object**	**Pictorial Object**
A1	0,19	0,07	0,16	0,08	0,15	0,09	0,18	0,07
A2	0,23	0,10	0,17	0,09	0,16	0,07	0,25	0,15
A3	0,14	0,17	0,10	0,08	0,17	0,11	0,18	0,09
A4	0,17	0,07	0,15	0,07	0,12	0,07	0,19	0,09
A5	0,25	0,10	0,12	0,17	0,25	0,21	0,18	0,09
A6	0,22	0,06	0,20	0,06	0,22	0,21	0,18	0,07
A7	0,23	0,07	0,13	0,11	0,16	0,16	0,18	0,08
A8	0,38	0,16	0,41	0,08	0,26	0,07	0,19	0,11
A9	0,29	0,07	0,31	0,07	0,19	0,20	0,15	0,06
A10	0,23	0,07	0,15	0,15	0,14	0,06	0,15	0,08
A11	0,19	0,06	0,14	0,16	0,12	0,06	0,15	0,06
A12	0,26	0,10	0,22	0,28	0,23	0,09	0,26	0,13
A13	0,55	0,28	0,07	0,06	0,10	0,03	0,08	0,28
A14	0,47	0,31	0,07	0,12	0,08	0,09	0,18	0,40
A15	0,32	0,33	0,06	0,07	0,40	0,09	0,19	0,30
A16	0,37	0,21	0,05	0,08	0,17	0,07	0,16	0,61
A17	0,30	0,25	0,10	0,13	0,20	0,09	0,06	0,28
A18	0,48	0,08	0,29	0,08	0,06	0,08	0,10	0,64
A19	1,10	0,14	0,37	0,12	0,10	0,10	0,18	0,82
A20	0,26	0,13	0,22	0,09	0,11	0,09	0,15	0,45
A21	0,41	0,08	0,22	0,07	0,03	0,07	0,10	0,24
A22	0,35	0,06	0,18	0,05	0,13	0,06	0,09	0,19
A23	0,33	0,08	0,19	0,07	0,14	0,07	0,19	0,28
A24	0,45	0,08	0,29	0,08	0,12	0,08	0,09	0,27
A25	0,33	0,11	0,21	0,07	0,02	0,06	0,04	0,64
A26	0,32	0,07	0,24	0,17	0,04	0,21	0,08	0,32
A27	0,24	0,07	0,16	0,06	0,04	0,06	0,08	0,26
A28	1,16	0,12	0,38	0,21	0,17	0,08	0,12	0,34
A29	0,11	0,05	0,12	0,03	0,03	0,02	0,06	0,11
A30	0,35	0,10	0,18	0,07	0,05	0,07	0,12	0,32
A31	0,24	0,08	0,12	0,28	0,07	0,16	0,06	0,24
A32	0,33	0,16	0,24	0,11	0,09	0,06	0,06	0,54
A33	0,19	0,05	0,14	0,16	0,04	0,06	0,06	0,40
$\overline{\sum}$	11,44	3,52	7,19	3,49	3,75	3,41	2,97	9,02

TABLE. II. COMPARISON OF TIME OF INTERACTION PAPER VS. DIGITAL

Paper	**Digital**
25,35	19,44

TABLE. III. THE RESULTS OF THE WILCOXON TEST

Test Statistics	
	Digital - Paper
Z	-2.976^{b}
Asymp. Sig. (2-tailed)	.003

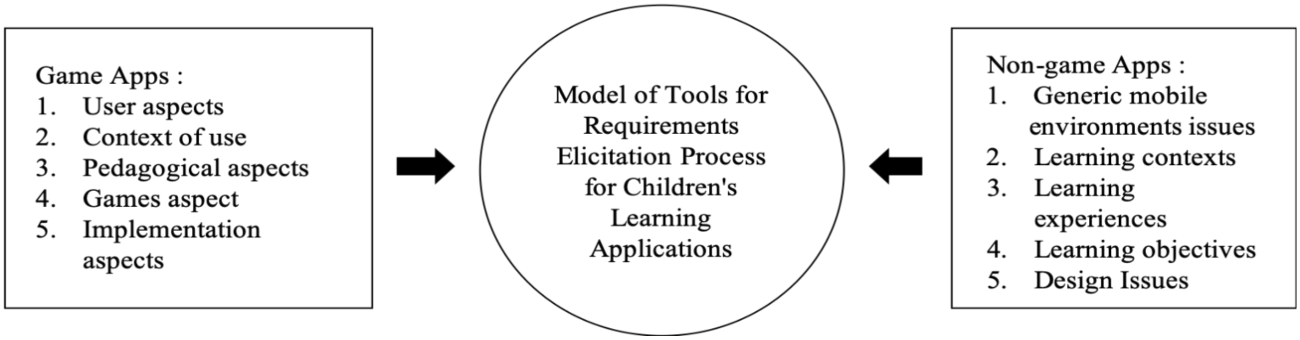


Fig. 3. Model of Tools for Requirements Elicitation Process for Children's Learning Application

The model of tools for requirements elicitation process for children's learning applications has several aspects, as shown in Fig. 3. Each aspect is distinguished based on the form of the application to be made. Application game aspects are user aspects, the context of use, pedagogical aspects, games aspect, and implementation aspects [26]. In the learning application, the type of games to be used are serious games. The reason for using this type is because serious games are tools that are considered useful in the learning process [27]. The material presented is distinguished based on thinking skills level in-game elicitation applications aimed at linking game mechanics that were commonly founded with learning mechanisms [25]. According to Piaget's, it was explained that children aged 6-8 years could have the ability to classify and understand ideas [28].

In non-game type applications, applications tend to be made in the form of simulations. Simulations can provide considerable learning potential because it is more effective and interactive [29] and can present material forms that convey the conditions of the situation in the real world [30]. The aspects are generic mobile environment issues, learning contexts, learning experiences, and learning objectives [31], as Fig. 3. The model of tools in Fig. 3 was implemented in the form of a mobile-based application. In the elicitation application that was built, all aspects of both types of applications will be accommodated. The selection of application types in the elicitation process was carried out at the beginning of the elicitation activity. The Requirements document can also be generated in this application. The aim is to facilitate and ease the elicitation team in verifying and validating requirements. Fig. 4 is a flowchart of the elicitation application that was built.

The VARK learning style implementation is applied to the elicitation applications that were built. The application is not only in the presentation of the material but also in the evaluation of the material. In presenting the material, children were given the opportunity to choose learning styles that suit their preferences. After the child listens to the material, then the evaluation is carried out. When evaluating given some questions related to the material have presented. Examples of problems can be seen in Table IV. The interaction of child answering questions was done according to the learning style chosen. Evaluation questions were presented according to the level of children's thinking skills. The flowchart of applying the VARK learning style can be seen in Fig. 5.

The implementation of elicitation tools that had been built in mobile-based applications. The application then tested on 32 children as respondents. Each child was asked to fill in each component by their preferences for each type of application. At the time of the presentation, the type of learning style was adjusted to the results of filling the VARK questionnaire given one week before the application testing process. The aim was made to facilitate the elicitation team in testing and assessing whether the learning style generated from the questionnaire is by the wishes of the child.

The results of filling out the questionnaire data can be seen in Table IV. From that data, a reliability test was performed using Cronbach's Alpha. The processing results obtained a value of 0.679, so it can be said that this questionnaire has reliability. Then do the processing of the results of data filling by respondents to assess the satisfaction of the application user, and the results obtained Fig. 6.

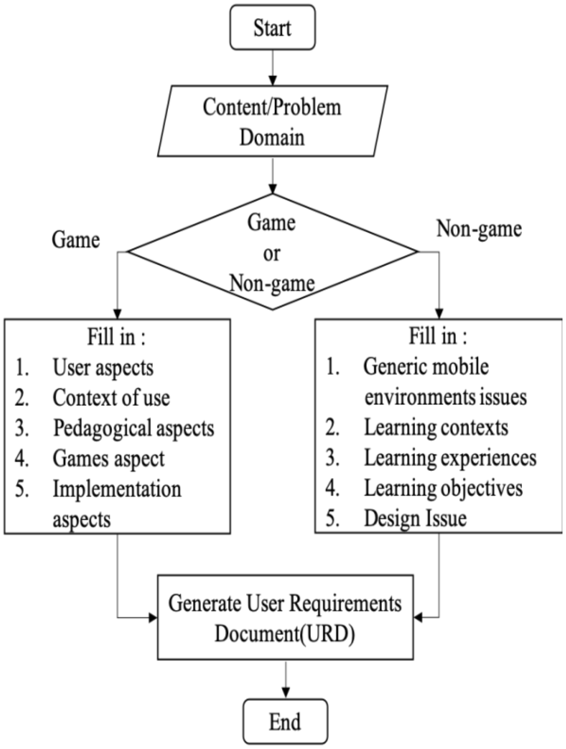


Fig. 4. Flowchart Model of Tools for Requirements Elicitation Process for Children's Learning Application

TABLE. IV. EXAMPLES OF QUESTIONS TO MEASURE THINKING SKILLS LEVEL FOR CASE NON-GAME APPLICATIONS

Thinking Skills Level	**Easy**	**Medium**	**Hard**
Retention (C1)	Choose three types of vegetables	Group the types of vegetables and fruits (5 each)	Group the types of green vegetables and fruits
Understandin g(C2)	How many vegetables are green?	If three vegetables are taken for cooking, how many vegetables now?	If each menu of dishes requires three types of vegetables, how many of each kind of vegetable needed to cook five menus of dishes?
Applying (C3)	If each menu of dishes requires three types of vegetables, how many of each kind of vegetables needed to cook five menus of dishes (The time given to answer is 30 seconds)	If the mother is going to make fruit juice as it is pictured, what fruit is needed? (3 juices were served in red, white, orange) (The time given to answer is 30 seconds)	There is a food and drink menu, do groupings of vegetables and fruits according to each menu (The time given to answer is 30 seconds)

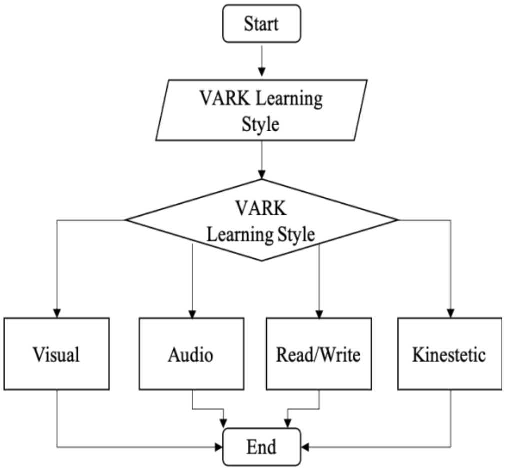


Fig. 5. Flowchart of Applying the VARK Learning Style.

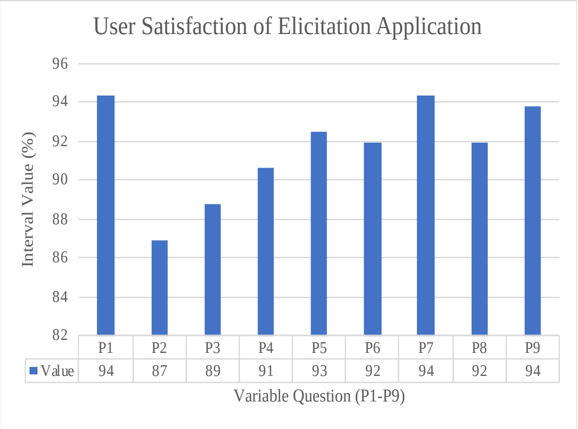


Fig. 6. Chart of user Satisfaction Assessment of Elicitation Applications.

Based on the results of processing, an interval analysis was performed to measure the satisfaction of the application. The results of the interval analysis can be concluded that each question (P1-P9) has a value >= 80% (strongly agree). In other words, the application of elicitation tools can be used by users well for each component of the question.

V. CONCLUSION

The conclusions that can be drawn from the results of this study are:

- The use of digital media was recommended making elicitation tools for children's learning software. This was evidenced by the non-parametric statistical tests using the Wilcoxon approach, which yields a value of 0.003. That means that there are differences in interactions between the use of paper and digital in terms of time.
- The model of tools elicitation that was built has a satisfaction level> = 80%. This means that children can express their desires on the learning application to be built. In other words, children's communication with the team can be done well.

Future work is to implement that elicitation tools model in real cases for children's learning applications

ACKNOWLEDGMENT

This research was supported by The Indonesia Endowment Fund for Education (Lembaga Pengelola Dana Pendidikan/LPDP).

REFERENCES

[1] Distanont and H. Haapasalo, "The Engagement between Knowledge Transfer and Requirements Engineering," Int. J. Manag. Knowl. Learn., vol. 1, no. 2, pp. 131–156, 2012.

[2] L. C. Ronoh, G. M. Muchiri, and F. Wabwoba, "Factors affecting requirements elicitation for heterogeneous users of information systems," Int. J. Comput. Sci. Eng. Technol. IJCSET), vol. 5, no. 3, pp. 35–39, 2015.

[3] B. Davey and C. Cope, "Requirements Elicitation - What's Missing?," Proc. 2008 InSITE Conf., vol. 5, 2008.

[4] A. Ferrari, P. Spoletini, and S. Gnesi, "Ambiguity and tacit knowledge in requirements elicitation interviews," Requir. Eng., vol. 21, no. 3, pp. 333–355, 2016.

[5] N. C. L. Hess, D. J. Carlson, J. D. Inder, E. Jesulola, J. R. Mcfarlane, and N. A. Smart, Clinically meaningful blood pressure reductions with low intensity isometric handgrip exercise. A randomized trial, vol. 65, no. 3. 2016.

[6] M. Ivanov and P. D. Werner, "Behavioral communication: Individual differences in communication style," Pers. Individ. Dif., vol. 49, no. 1, pp. 19–23, 2010.

[7] A. De Lucia and A. Qusef, "Requirements engineering in agile software development," J. Emerg. Technol. Web Intell., vol. 2, no. 3, pp. 212–220, 2010.

[8] A. Davis et al., "Identifying and Measuring Quality in a Software Requirements Specification," in Software Metrics Symposium, 1993. Proceedings., First International, 1993, pp. 141–152.

[9] P. D. Chatzoglou and L. A. Macaulay, "Requirements capture and IS methodologies," Inf. Syst. J., vol. 6, no. 3, pp. 209–225, 1996.

[10] A. M. Aranda, O. Dieste, and N. Juristo, "Effect of Domain Knowledge on Elicitation Effectiveness: An Internally Replicated Controlled Experiment," IEEE Trans. Softw. Eng., vol. 42, no. 5, pp. 427–451, 2016.

[11] H. F. Hofmann and F. Lehner, “Requirements engineering as a success factor in software projects,” IEEE Softw., vol. 18, no. 4, pp. 58–66, 2001.

[12] C. K. Gonzales and G. Leroy, “Eliciting user requirements using Appreciative inquiry,” Empir. Softw. Eng., vol. 16, no. 6, pp. 733–772, 2011.

[13] K. Siau and X. Tan, “Using cognitive mapping techniques to supplement UML and UP in information requirements determination,” J. Comput. Inf. Syst., vol. 46, no. 5 SPEC. ISS., pp. 59–66, 2006.

[14] H. Al-Zawahreh and K. Almakadmeh, “Procedural model of requirements elicitation techniques,” ACM Int. Conf. Proceeding Ser., vol. 23-25-Nove, 2015.

[15] N. Iivari, “‘Representing the User’ in software development-a cultural analysis of usability work in the product development context,” Interact. Comput., vol. 18, no. 4, pp. 635–664, 2006.

[16] T. Nousiainen, “Children’s Involvement in the Design of Game-Based Learning Environments Cases Talarius and Virtual Peatland,” Des. Use Serious Games, vol. 37, pp. 49–66, 2009.

[17] S. Livingstone, “Reframing media effects in terms of children’s rights in the digital age,” J. Child. Media, vol. 10, no. 1, pp. 4–12, 2016.

[18] J. W. Santrock, Life-Span Development. The McGraw-Hill Companies, 2012.

[19] J. M. Carrillo De Gea, J. Nicolás, J. L. Fernández Alemán, A. Toval, C. Ebert, and A. Vizcaíno, “Requirements engineering tools: Capabilities, survey and assessment,” Inf. Softw. Technol., vol. 54, no. 10, pp. 1142–1157, 2012.

[20] M. K. Sabariah, P. I. Santosa, and R. Ferdiana, “Selecting elicitation technique on requirements elicitation process: A case study on education application for children,” in IOP Conference Series: Materials Science and Engineering, 2018, vol. 434, no. 1.

[21] J. E. Helmreich, “Statistics: An Introduction Using R (2nd Edition) ,” J. Stat. Softw., vol. 67, no. Book Review 5, 2015.

[22] S. M. Fisch, S. Damashek, and F. Aladé, “Designing media for cross-platform learning: Developing models for production and instructional design,” J. Child. Media, vol. 10, no. 2, pp. 238–247, 2016.

[23] J. Beatty and A. Chen, Visual Models for Software Requirements: An RML® Handbook. 2012.

[24] S. Cano, C. Collazos, H. M. Fardoun, and D. M. Alghazzawi, “Model Based on Learning Needs of Children,” Int. Conf. Soc. Comput. Soc. Media, vol. 3, pp. 324–334, 2016.

[25] S. Arnab et al., “Mapping learning and game mechanics for serious games analysis,” Br. J. Educ. Technol., vol. 46, no. 2, pp. 391–411, 2015.

[26] O. De Troyer and E. Janssens, “Supporting the requirement analysis phase for the development of serious games for children,” Int. J. Child-Computer Interact., vol. 2, no. 2, pp. 76–84, 2014.

[27] A. Slimani, O. B. Yedri, F. Elouaai, and M. Bouhorma, “Towards a design approach for serious games,” Int. J. Knowl. Learn., vol. 11, no. 1, pp. 58–81, 2016.

[28] R. D. Vatavu, G. Cramariuc, and D. M. Schipor, “Touch interaction for children aged 3 to 6 years: Experimental findings and relationship to motor skills,” Int. J. Hum. Comput. Stud., vol. 74, pp. 54–76, 2015.

[29] L. Sha, C. K. Looi, W. Chen, P. Seow, and L. H. Wong, “Recognizing and measuring self-regulated learning in a mobile learning environment,” Comput. Human Behav., vol. 28, no. 2, pp. 718–728, 2012.

[30] M. E. Gredler, Games and Simulations and Their Relationships to Learning. 2004.

[31] D. Parsons, H. Ryu, and M. Cranshaw, “A design requirements framework for mobile learning environments,” J. Comput., vol. 2, no. 4, pp. 1–8, 2007.

Intelligent System for Price Premium Prediction in Online Auctions

Mofareah Bin Mohamed[1], Mahmoud Kamel[2]
Information Systems Department
Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—The use of data mining techniques in the field of auctions has attracted considerable interest from the research community. In auctions, the users try to achieve the highest gain and avoid loss as much as possible. Therefore, data mining techniques can be implemented in the auctioning domain to develop an intelligent method that can be used by the users in online auctions. However, determining the factors that affect the result of an auction, especially the initial price, is critical. In addition, the intelligent system must be established based on clean data to ensure the accuracy of the results. In this paper, we propose an intelligent system (classifier) to predict the initial price of auctions. The proposed system uses the double smoothing method (DSM) for data cleaning in terms of preprocessing. This system is implemented on a data set collected from the eBay website and cleaned using the proposed DSM. In the training phase, the CART technique is employed for the classifier construction. Compared to similar techniques, the proposed system exhibits a better performance in terms of the accuracy and robustness against noisy data, as determined using ROC curves.**

Keywords—Classification; auction; CART; training; testing; preprocessing; noise; outlier; DSM

I. Introduction

Importance of the eBay website and auctions. The eBay website is one of the most important websites in the business area, and it can be considered as a leader among e-commerce and internet sites. eBay is used to sell and buy goods and products online or provide services worldwide. The importance of eBay is a result of the financial benefits generated by more than 168 million active buyers and more than 20 million active sellers, with billions of dollars of transactions occurring on the site [1]. By participating in online auctions, people can directly obtain and purchase the items that they desire, without the hassle/risk of traveling.

Motivation and problem statement. The final bid price of an auction is tightly coupled with the influencing factors. Although these factors were addressed in many previous works, the price premium was not considered. In the context of online auctions, the price premium is defined as “the monetary amount above the average price received by multiple sellers for a certain matching product” [2]. In addition, changes in the price premiums can indicate product shortages, excess inventories, or other changes in the relationship between the supply and demand. This aspect thus affects the gain earned by businesspeople. Therefore, the corresponding research question pertains to the determination or prediction of the factors that are related to and affect the price premium, thereby enabling businessmen to achieve the best gain and avoid loss by correctly estimating the initial price of the product. In addition to determining the factors, cleaning the data and eliminating the noise and outliers are critical to ensure accurate results.

The use of artificial intelligence in various domains has received considerable attention from the research community. In particular, by using an intelligent machine that takes some of the influencing factors as inputs, businesspeople can predict the price premium. In this context, the contributions of this work are as follows:

- A novel preprocessing method for data cleaning, known as the double smoothing method (DSM), is proposed. This method uses the binning method to process the noisy data, and the clustering-based technique is later used to filter the outliers.
- The CART technique is used to produce the decision rules for sellers in online auctions. CART is a type of decision tree, which acts as a classifier, and it can be used in the classification process to address categorical data as a final decision (in this work, reaching or not reaching the price premium). Furthermore, this technique can be used for regression to deal with continuous data [3].
- The variables that exert the most considerable effects on the auction outcomes are examined using the CART technique.
- Extensive experiments are performed on real data driven from the eBay website to evaluate the proposed approach against other similar approaches.

The remaining work is structured as follows: Section II presents the related work, followed by the description of the proposed artificial system and approach in Section III. Section IV describes the metrics used for the evaluation, and the subsequent section presents the experimental results and evaluations. Finally, the work is concluded in Section VI.

II. Related Work

In this section, first, we present an overview of internet auctions and later describe the decision tree induction techniques. Finally, we explore some works related to the conduction of auctions.

A. *Internet Auctions*

The development of the internet led to a revolution in the field of auctions, which were conducted only physically in the past. By using the internet, auctions could be conducted via emails or discussion lists [4]. At the end of the 1990s and beginning of the 2000s, the number of auction sites was estimated to be 200 [5]. Subsequently, competing sites were generated to manage auctions by famous companies, such as QXL.com in Europe, Taobao.com in Asia, and MercadoLibre in Latin America. Recently, the website of the eBay company has become the most famous site used to conduct auctions.

B. *Decision Tree Induction Techniques*

Decision trees can be considered as a reflection of the rules used for the classification in artificial intelligence research. Such trees visually represent the rules in the form of nested if–then statements [6]. Various algorithms are utilized to form decision trees, such as the CART, QUEST, ID3 and CHAID. These algorithms differ in terms of the mechanism used to determine the root of the tree and subsequently form the other branches. In addition, these techniques differ depending on the data type involved in the manipulation. For example, ID3 and CHAID can manipulate only categorical data, while CART can manipulate both categorical and continuous data [7].

The stochastic differential equation (SDE) has been used to model eBay prices [8]. In this study, the authors utilized the SDE to represent the price velocity and accelerator. In the experiments, as the database, the authors used 63 training samples and 30 testing samples from the auctions of the Microsoft X-box gaming system. Subsequently, by performing a differential analysis, the authors extracted the features and collected the results. Most importantly, it was indicated that the use of the SDE is more suitable for this task than the ordinary differential equation (ODE) approach [9].

The authors of the work [10] previously proposed a price prediction and insurance service for online auctions. The final objective of this service was to guarantee the minimum end price for the sellers in auctions such as those conducted on eBay. It was concluded that auctions with a reserve price option lead to a worse price at the end, which is the underlying reason for why the use of the price insurance service is desirable. For data gathering, a crawler was employed to collect data from the eBay website for two months. The features extracted to build the database were related to the seller, items, and the auction. By using multiple classification regressions algorithms, the final classifier was generated to classify the new data.

Dass et al. proposed a dynamic price forecasting method [11]. The features that distinguish this work are as follows: (1) the technique manipulates the same product that is involved in multiple auctions; (2) the price dynamics are considered, and the static data such as the initial price and seller reputation are ignored; and (3) the source of the price dynamics is manipulated in the context of the buyer's competitions. The drawback of ignoring such static data is that the initial price can be variable, and it depends on the seller opinion rather than eBay's policy, which in turn increases the bidding price, especially at the first stage of the auction.

A neural network-based approach was proposed [12] to solve nonparametric price prediction models. The key idea was to map the nonlinear data and approximate the end price regardless of any assumptions, by adjusting the weights of the inputs of the neural network.

Gregg et al. proposed an intelligent recommendation system, which can act as an adviser to the users for price prediction [13]. Under the time performance term and to make bidding decisions within a short amount of time, this system targets the search for bargain processes. The key idea is to present the users with relevant information such as the current bid and a recommended price based on the recently closed auctions. This system was enhanced in another study [14], primarily by linking the price prediction process with various features of the auction, such as the feedback rating and item description.

III. Proposed Artificial System Architecture

In this section, we describe the system architecture and present the details regarding the data set used for the training and testing phases.

A. *Artificial System Architecture*

The general system architecture consists of three main components, as illustrated in Fig. 1.

As shown in Fig. 1, the first component of the system is the database, which is represented by tables that contain data. The second component is the classifier, which needs to train considering the data collected and stored in the database. The trained classifier follows certain rules, which may be complex. Therefore, the third component is responsible for constructing and pruning the decision tree that the classifier later uses to make the final decision (i.e., classifying a new record or unknown data).

B. *Used Data Set*

Table I summarizes the data set used in this work, on which the classifier is trained and tested.

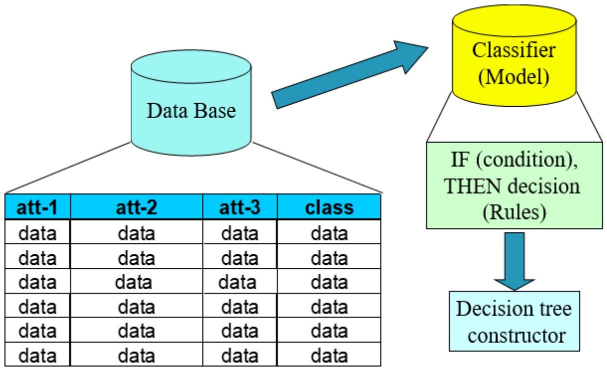


Fig. 1. Proposed System Architecture.

TABLE. I. Used Data Set

Location	Period	Item	No. of records
eBay U.S. website [15]	1 month	Palm Pilot M515 PDA	10,000

The eBay website is used for the data collection for the following reasons: First, this website includes real data, which is preferable when conducting practical experiments [16]. Second, the website is considered as a continuous resource of data by the sellers because it often motivates the users to compete at any time and in any location [17]. Finally, the mechanism employed by eBay is favorable to the auctioneers in most cases [18].

Data preprocessing. The data collected from the eBay website represent real data. In data mining, real data are often considered dirty, as they may be incomplete, noisy, inconsistent, missed, or including outliers. Such dirty data (due to instrument faults, human or computer errors, or transmission errors) negatively affects the quality of the intelligent system and its outcomes in terms of the accuracy [19, 27]. In addition, many issues should be taken into consideration during the preprocessing phase, such as ensuring data privacy [28, 29, 30, 31]. Moreover, enhancing the performance using high performance computing techniques [32] as well as ensuring the security of the data using some hiding or blurring techniques is required [33], or employing agent based software technology for solving transmission challenge problem [34]. The previous issues do not be taken into consideration in this work, and they will be manipulated as a future work.

To solve this problem and clean the data, a double smoothing method (DSM) is utilized, in which the binning method is followed by a clustering based technique. The binning method is used for data smoothing to eliminate noisy data [20], and it consists of two main steps: (1) sorting the data and partitioning them into (equal frequency) bins; and (2) smoothing the data (the boundary based method is used in this study [21]). The goal of the clustering method is to detect and eliminate the outliers, which are considered as the most negative type of noises that can be located within the data. In the context of this work, the outliers refer to the extremely low (or high) values of the attributes used for constructing the database [22].

After cleaning and smoothing the data, the curse of dimensionality problem may arise. Because the analysis of complex data (that include many attributes or dimensions) on the complete data set may require a considerable amount of time, the dimensionality of the data must be reduced. In this work, we rely on the feature selection method to reduce the dimensionality [23]. After this process, the obtained database, as shown in Fig. 2, can be used to train the classifier.

The features selected to reduce the dimensionality are those that have the most considerable impact on the auction decision. Such factors include the shipping cost, reputation (expressed by rating), initial bid price, and auction ending time. These features are considered as the variables that determine the final price in the context of the auction process. Statistics operations are applied to the auction data to obtain the descriptive statistics, that is, the mean and standard deviation for continuous or real (float values) data variables and the frequencies of the categorical data. Table II summarizes the descriptive statistics obtained in this work.

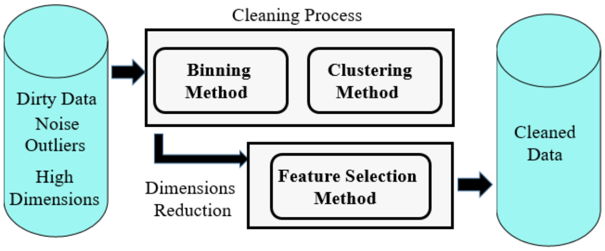


Fig. 2. Cleaning and Data Reduction.

TABLE. II. DESCRIPTIVE STATISTICS

	Palm Pilot M515 PDA	
	Mean	***St. dev***
Criterion dependent variable ***Final Bid Price (FBP)***	229.4409	21.9659
Independent numerical variables		
Initial Bid Price	78.1442	92.1557
Auction Duration	5.5882	1.74
Number of Bids	17.3706	11.2954
Independent categorical variables		
	Frequency	
Auction Ending Time: 1) Weekday Morning 2) Weekday Afternoon 3) Weekend Morning 4) Weekend Afternoon	 105 129 60 46	

C. Model Construction (Classifier)

To create the classifier, the cross-validation method is used, which is a common tool in data mining. This approach consists of two main stages, namely, the training stage and testing stage [24]. The final goal of the training stage is to construct the classifier by training it on the dataset, while the objective of the second stage is to estimate the performance of the classifier in terms of the accuracy. The cross-validation method involves two main steps: (1) randomly partitioning the data into k mutually exclusive subsets, with all the subsets having an approximately equal size; and (2) at the i^{th} iteration, using D_i as the test set and other sets as the training set. Fig. 3 illustrates the process flow of the cross-validation method.

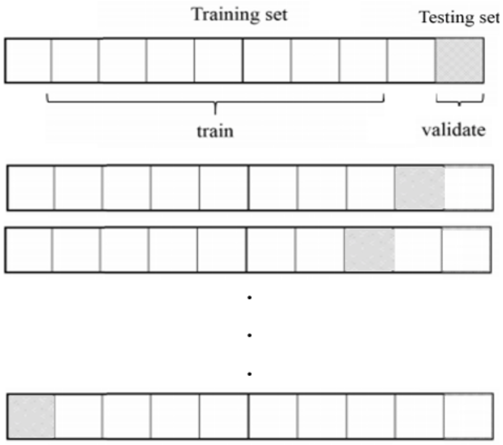


Fig. 3. Cross-Validation Method, k=10.

The classifier follows certain rules in the process of defining the initial price to decide if a user should continue in the auction or not. In other words, the class that the classifier predicts is the initial price based on the following rules that are formed using (and, or) operators located among the predefined features. The rules are considered as the training space and used to form a decision tree.

Algorithm 1 shows the steps of the cross-validation method.

Algorithm 1: Cross-validation

```
Input: training dataset
Output: performance estimation of models
initialization;
foreach Parameter set do
    Load the classifier with new parameters
    Split into K validation sets
    for $k \in K$ do
        Train on training set
        Test on validation set
        Determine the performance
    end
    Calculate the performance mean in the K validation
end
```

D. *Construction of the Decision Tree*

The classifier follows certain rules in the process of defining the initial price to decide if a user should continue in the auction or not. In other words, the class that the classifier predicts is the initial price based on the following rules that are formed using (and, or) operators located among the predefined features. The rules are considered as the training space and used to form a decision tree.

The CART algorithm, which involves a nonparametric procedure, is employed to create the optimal decision tree. The main advantage of CART is that it supports certain data types in the classification and real or continues data types in the regression. The findings obtained using the CART are simple to understand and visualize. The strategy followed by the CART to construct the decision tree can be summarized as follows:

1) Select features or variable. In this step, the features extracted to achieve the dimensionality reduction are used as the variables for the CART algorithm.

2) Determine the splitting condition, that is, determine the best selected features as the root of the decision tree.

3) Determine the stopping criteria, which indicates the completion of the decision tree. In this work, the stop condition is achieved when no more data are available in the data set.

4) Perform pruning, which is aimed at avoiding the overfitting problem. In this work, this problem is avoided as the most suitable features are selected as variables for the CART algorithm and double data cleaning is performed against noisy data and outliers.

IV. Evaluation Metrics

In this work, two main performance metrics are utilized for the evaluation: the confusion matrix, and the ROC curve, which is inspired from the confusion matrix.

A. *Confusion Matrix*

In general, the confusion matrix is a useful tool for analyzing how well a classifier can recognize the tuples of different classes. The confusion matrix is formed considering the following terms [25]:

1) True positives (TP): positive tuples that are correctly labeled by the classifier.

2) True negatives (TN): negative tuples that are correctly labeled by the classifier.

3) False positives (FP): negative tuples that are incorrectly labeled as positive.

4) False negatives (FN): positive tuples that are mislabeled as negative.

Table III shows the confusion matrix in terms of the TP, FN, FP, and TN.

Depending on the confusion matrix, the accuracy of a given classifier can be calculated by considering the recognition rate, which is the percentage of the test set tuples that are correctly classified. The accuracy can be obtained using the following formula:

$$accuracy = \frac{(TP+TN)}{number\ of\ all\ records} \quad (1)$$

Accuracy based valuation. In this context, a higher accuracy corresponds to a better classifier output. The maximum value of the accuracy metric is 1 (or 100%), which is achieved when the classifier classifies the data correctly without any error in the classification process.

B. *ROC Curve*

Receiver operating characteristic (ROC) curves are used to enable the visual comparison of different classification models. These curves indicate the balance between the true and false positive rates, and the area under the ROC curve denotes the accuracy of the classifier [26].

ROC based evaluation. In this context, a model representing a line closer to the diagonal line (i.e., the closer the area is to 0.5) is a less accurate model.

TABLE. III. Confusion Matrix

Actual class (Predicted class)	**C1**	**¬ C1**
C1	True Positives (TP)	False Negatives (FN)
¬ C1	False Positives (FP)	True Negatives (TN)

V. Experimental Results and Evaluations

The proposed approach is implemented using the MATLAB programming language. The system is executed on a laptop with the following configuration: Genuine Intel(R) 2.4 GHz PC with 4.00 G RAM, running Microsoft Windows 7 Ultimate. The proposed system is compared with the recommendation system for price prediction (RSPP) that was presented in [14].

1) Evaluation based on the confusion matrix: In this context, the same data set size is used to enable a fair comparison. C1 refers to the suitable predicted initial price, while ¬ C1 refers to the unsuitable predicted initial price. These aspects are represented as C1=yes and ¬ C1= no.

Table IV lists the values of the confusion matrix for both the proposed approach and the RSPP.

$$accuracy\ (proposed\ system) = \frac{9542}{10000} \approx 95\%$$

$$accuracy\ (RSPP\ system) = \frac{7454}{10000} \approx 75\%$$

Discussion. The proposed system outperforms the RSPP system in terms of the accuracy. This finding can be attributed to the training phase. In the proposed system, the cross-validation is performed 10 times, which means that the classifier is trained on the complete data set. In other words, in each run, a part of the data set is used a training set, which leads to comprehensive training, thereby providing the classifier with more alternatives to deal with new data. In the RSPP system, a holdout method is used, which divides the complete data set into two main data sets (training and testing). However, the training and testing set constitute 80% of the original data set, respectively. Since the training set is employed to construct the model (i.e., the classifier), the time spent in the training phase is considerable smaller compared to that in the proposed system. It is known that a higher training time leads to more accurate outputs.

2) Evaluation based on ROC curves: In this context, we use the same data set size under the same conditions considered in the previous comparison. In addition, we evaluate the systems involved in the comparison in terms of the robustness. The robustness refers to the ability of a classifier to provide correct predictions in the case of noisy data or data with outliers.

TABLE. IV. Confusion Matrix

	Actual class (Predicted class)	**Yes**	**No**	**Total**
Proposed system	Yes	6954	46	7000
	No	412	2588	3000
	Total	7366	2634	10000
RSPP system	Yes	4954	96	5050
	No	2450	2500	4950
	Total	7404	2596	10000

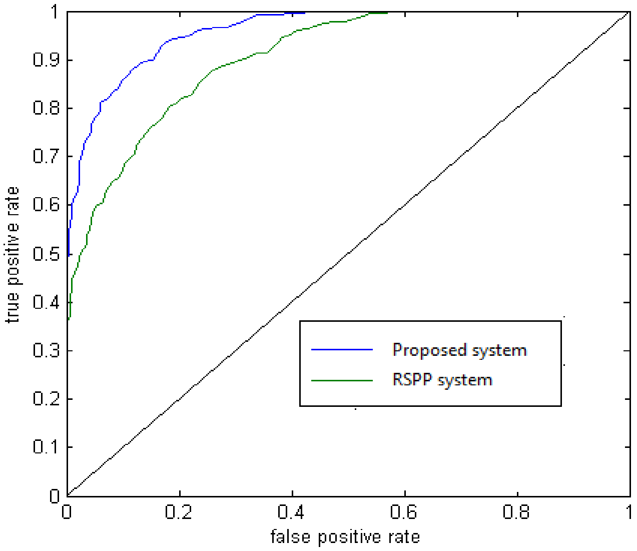


Fig. 4. ROC Curves.

Fig. 4 shows the ROC curves for both the proposed system and the RSPP system.

Discussion. Fig. 4 illustrates that the proposed system performs better than the RSPP system under noisy data (particularly, in the presence of outliers). In general, the accuracy of a classifier is negatively affected by outliers because they lead to dramatic decreased (or increased) values that are reflected as low accuracy in the classification process. The low accuracy is a normal result of certain rules generated (which are suitable only) for outliers. However, the proposed system can address both noisy data and outliers effectively because the preprocessing step is performed before training the classifier (i.e., the DSM). The noisy data added to the cleaned data for robustness testing are filtered using the binning method. Moreover, the outliers inserted within the original data are groped and deleted using the clustering-based method. Consequently, the abnormal data are filtered before training the classifier in the proposed system, which is reflected in a high classification accuracy. In the RSPP system, bag of words (BOW) as well as frequency-based methods are used to preprocess the data. However, a certain threshold of frequency is used to determine if a given value corresponds to a noisy data point or outlier. Therefore, many values are used in the training phase. The presence of such abnormal data (that are not filtered and contribute to the classifier construction) leads to the lower classification accuracy compared to that of the proposed system.

VI. Conclusion

The final bid price of the auction is tightly coupled with the factors influencing the price. Among these factors, the initial price plays a vital role in the decision making process conducted by the user for being involved or not in an auction. To realize effective decision making and help users, artificial intelligence techniques can be employed. The process of development of an intelligent system involves two main phases, namely, training and testing. However, preprocessing the data used in the training phase for the model construction is critical to ensure accurate results. In this work, we employ the double smoothing method (DSM) for data cleaning. The data are cleaned by subjecting them to two processes, namely, (1) the binning method which is responsible for noisy data elimination; and (2) the clustering-based method, which is responsible for outlier detecting and deletion. The classifier is built based on the cleaned data by using the cross-validation method (with k=10). The data that the classifier trains on are collected from the eBay website and arranged in a database of (1000) cleaned records. The decision tree, which contains the rules that the classifier follows in the process of classification, is formed using the CART algorithm as it can deal with numerical, continuous and categorical data. A confusion matrix and ROC curves are employed to evaluate the proposed system against similar systems. The results show that the proposed system achieved an accuracy of 95% compared to that of 75% for an existing system. This result is supported by the ROC curves, which indicate that the proposed system exhibits a better accuracy and robustness against noisy data and outliers.

In future work, we intend to enhance the proposed system to achieve a higher accuracy and ensure the privacy protection of the manipulated data. In addition, different data set can be used to prove scalability of this proposed work.

References

[1] Mishra, Miltan Kumar. "Why is eBay the Most Successful Online Auction?." Global Journal of Management And Business Research 10.9 (2010).

[2] Ba, S., & Pavlou, P. A. T2002. Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. MIS quarterly, 247-248.

[3] Breiman, L., Friedman, J., Olshen, R., & Stone, C. 1984. Classification and regression trees. Belmont, CA: Wadsworth International Group.

[4] Lucking-Reiley, David. "Auctions on the Internet: What’s being auctioned, and how?." The journal of industrial economics 48.3 (2000): 227-252.

[5] Crockett, R. O. "Going, going… richer." Business Week 3659 (1999): EB16.

[6] Kim, Jong Woo, et al. "Application of decision-tree induction techniques to personalized advertisements on internet storefronts." International Journal of Electronic Commerce 5.3 (2001): 45-62.

[7] Zanakis, Stelios H., and Irma Becerra-Fernandez. "Competitiveness of nations: A knowledge discovery examination." European journal of operational research 166.1 (2005): 185-211.

[8] Liu, W. W., Liu, A., & Chan, G. H. 2018. Modeling eBay Price Using Stochastic Differential Equations. Journal of Forecasting.

[9] Hsu, S. B. 2013. Ordinary differential equations with applications (Vol. 21). World Scientific Publishing Company.

[10] Ghani, R. 2005. Price prediction and insurance for online auctions. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (pp. 411-418). ACM.

[11] Dass, M., Jank, W., &Shmueli, G. 2010. Dynamic price forecasting in simultaneous online art auctions. In Marketing Intelligent Systems using Soft Computing (pp. 417-445). Springer, Berlin, Heidelberg.

[12] Ince, H., &Trafalis, T. B. 2007. Kernel principal component analysis and support vector machines for stock price prediction. IIE Transactions, 39(6), 629-637.

[13] D.G. Gregg and S. Walczak. Auction Advisor: an Agent-based Online Auction Decision Support System. Decision Support Systems, 41(2):449–471, 2006.

[14] Van Heijst, Dennis, Rob Potharst, and Michiel van Wezel. "A support system for predicting eBay end prices." Decision Support Systems 44.4 (2008): 970-982.

[15] Baker, J., & Song, J. 2008. Exploring decision rules for sellers in business-to-consumer (b2c) internet auctions. International Journal of E-Business Research (IJEBR), 4(1), 1-21.

[16] Levin, Dan. "Demand reduction in multi-unit auctions: evidence from a sportscard field experiment: comment." American Economic Review 95.1 (2005): 467-471.

[17] Wingfield, Nick. "Corporate sellers put the online auctioneer on even faster track." Wall Street Journal 237.107 (2001): A1.

[18] Turban, Efraim, et al. "Business-to-Business E-Commerce." Electronic Commerce 2018. Springer, Cham, 2018. 123-166.

[19] Laboreiro, Gustavo Alexandre Teixeira. "Noise reduction and normalization of microblogging messages." (2018).

[20] Hariharakrishnan, Jayaram, S. Mohanavalli, and KB Sundhara Kumar. "Survey of pre-processing techniques for mining big data." 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP). IEEE, 2017.

[21] Siahaan, Andysah Putera Utama. "Quality Assurance in Knowledge Data Warehouse." (2017).

[22] Zimek, Arthur, and Peter Filzmoser. "There and back again: Outlier detection between statistical reasoning and data mining algorithms." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8.6 (2018): e1280.

[23] Novaković, Jasmina. "Toward optimal feature selection using ranking methods and classification algorithms." Yugoslav Journal of Operations Research 21.1 (2016).

[24] Varoquaux, Gaël. "Cross-validation failure: small sample sizes lead to large error bars." Neuroimage 180 (2018): 68-77.

[25] Düntsch, Ivo, and Günther Gediga. "Confusion matrices and rough set data analysis." Journal of Physics: Conference Series. Vol. 1229. No. 1. IOP Publishing, 2019.

[26] Obuchowski, Nancy A., and Jennifer A. Bullen. "Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine." Physics in Medicine & Biology 63.7 (2018): 07TR01.

[27] Alrahhal, Mohamad Shady, and Adnan Abi Sen. "Data mining, big data, and artificial intelligence: An overview, challenges, and research questions." (2018).

[28] Alrahhal, Mohamad Shady, Maher Khemakhem, and Kamal Jambi. "Agent-Based System for Efficient kNN Query Processing with Comprehensive Privacy Protection." International Journal Of Advanced Computer Science And ApplicationS 9.1 (2018): 52-66.

[29] Alrahhal, Mohamad Shady, Maher Khemakhem, and Kamal Jambi. "A Survey On Privacy Of Location-Based Services: Classification, Inference Attacks, And Challenges." Journal of Theoretical & Applied Information Technology 95.24 (2017).

[30] Alrahhal, Mohamad Shady, et al. "AES-route server model for location based services in road networks." Int. J. Adv. Comput. Sci. Appl 8.8 (2017): 361-368.

[31] Alrahhal, Mohamad Shady, Maher Khemekhem, and Kamal Jambi. "Achieving load balancing between privacy protection level and power consumption in location based services." (2018).

[32] Fouz, Fadi, et al. "Optimizing Communication And Cooling Costs In Hpc Data Center." Journal of Theoretical and Applied Information Technology 85.2 (2016): 112.

[33] Al-Rahal, M. Shady, Adnan Abi Sen, and Abdullah Ahmad Basuhil. "High level security based steganoraphy in image and audio files." Journal of theoretical and applied information technology 87.1 (2016): 29.

[34] Bandar Alluhaybi, Mohamad Shady Alrahhal, Ahmed Alzhrani and Vijey Thayananthan, “A Survey: Agent-based Software Technology Under the Eyes of Cyber Security, Security Controls, Attacks and Challenges” International Journal of Advanced Computer Science and Applications(IJACSA), 10(8), 2019. http://dx.doi.org/10.14569/IJACSA.2019.0100828.

Applying Social-Gamification for Interactive Learning in Tuberculosis Education

Dhana Sudana[1], Andi W.R. Emanuel[2*], Suyoto[3]
Universitas Atma Jaya Yogyakarta
Sleman, Indonesia 55281

Ardorisye S. Fornia[4]
PKU Muhamadiyah Yogyakarta Hospital
Yogyakarta, Indonesia 55122

***Abstract*—There are several methods of education for tuberculosis, and one of them is through the DOTS (Direct Observed Treatment Shortcourse) program. The management of tuberculosis education through the DOTS program is performed in clinics and hospitals only to patients and their families. The purpose of this study is to describe the development and testing of a prototype (social-game education) for interactive education for tuberculosis patients in particular and the general public. The data collection process is through direct observation of tuberculosis patients and health professionals (doctors, nurses, and DOTS health workers). Challenge the game in the prototype was provided with content that contained tuberculosis information that had been previously validated by a specialist. In addition to tuberculosis information as the main content, two important elements are making up this prototype, which are gamification and social media elements. In the game elements, this study adopted elements of the leaderboard, badge/achievement, challenge, and level. As for the third element, social media includes likes, comments, and shares. Prototype application testing was conducted on two participant groups (N = 48) consisting of 23 tuberculosis patients and 25 random participants. By using the user experience questionnaire (UEQ) technique, this research focuses on identifying the user's motivation in capturing compositional information as well as the clarity of the prototype. With a confidence interval of 5% (p = 0.05) per scale. The results indicate that participants have a high level of motivation towards the prototype; this is seen in the rating scale of stimulation with an average of 1.578. Likewise the effectiveness level of information in the rating scale of perspicuity has a mean of 1.224 also has a rate that is quite effective.**

***Keywords*—Education; gamification; mobile application; social-media; tuberculosis**

I. Introduction

Based on the WHO report in 2017, Indonesia is one of the countries with the highest number of tuberculosis sufferers in the world. Even tuberculosis infection is still the number one cause of death in Indonesia. The tuberculosis incidence rate in Indonesia is 395 per 100,000 population, and the death rate is 40 people per 100,000 per year [1]. Preventing transmission by optimizing tuberculosis case finding and healing is a program through DOTS (Direct Observed Treatment Shortcourse) [2]. Through the DOTS program [3][4], the method of prevention by detecting tuberculosis transmission among family members and those closest to the patient includes maximizing the tuberculosis education program. Actually, besides being able to support the prevention of tuberculosis, broad education about the disease can also reduce the unfortunate stigma of tuberculosis patients. In an all-digital era with a large community, mobile access can be an easy way in tuberculosis education widely, including through games and social media. Education through gamification methods [5] and social media methods [6] are proven to increase motivation and learning achievement.

Today's gaming applications continue to develop primarily as a means of education; even many gaming applications adopt elements of social media or social - gamification. The combination of social media elements in gamification gives the individual creating a profile or person and interacting with other users in each gameplay. Many educational institutions use the aspects of social media and games as a means of learning to motivate students and teachers alike [7]. Likewise, in the field of health, some elements of games and social media are used in various applications to help motivate patients undergoing therapy as well as stimulating a healthy lifestyle [8].

Based on the level of effectiveness of the role of games and social media in the world of education, this research focuses on the design of social-gamification content as a means of tuberculosis information and knowledge. Lung specialist doctors validate the social-games content to test the truth about tuberculosis educational content. Giving an achievement in the form of a badge or points to the player can motivate users in learning and explore the deeper aspects of education [9]. In contrast, the adoption of the elements like likes and comments contained in social media gives users a share of achievements to friends who are on the friend’s list.

In this study, we designed a prototype of tuberculosis disease learning applications. Our specifics are adopting social media and utilizing elements of gamification. By using the elements of social media, such as sharing, liking, and commenting, and utilizing the two elements of gamification elements, namely the mechanical and dynamic elements. The focus of the evaluation is to identify two elements of effectiveness, namely participant motivation and understanding of prototype content, whose ultimate goal is education. This social-gamification prototype was tested on participants using the user experience questionnaire (UEQ) method [10]. We differentiate between the two participant populations to compare the effectiveness of TB education learning. In this study, we also tried to identify user reactions by analyzing participants' opinions.

*Corresponding Author

II. Related Work

There are several studies on social-gamification for tuberculosis education. A survey on tuberculosis education explains, in general, tuberculosis disease education uses a conventional approach through the DOTS (Direct Observed Treatment Shortcourse) program. In the DOTS program, the patient and the patient's family are given oral education to motivate tuberculosis patients and cope with drug withdrawal behavior [11]. But in the study of Lam, et al. in 2018 explained that education and care for tuberculosis patients could be done remotely with Video-DOTS [12]. Ilya et al. also demonstrate that distance education through online games provides more effective benefits because it can find out study time, number of social relationships, and get bonus game points [13].

Several previous studies have explained that more and more educational methods in the prevention of disease and health care through gamification methods [14][15]. Research by Hursen et al. in 2019, about the purpose of gamification in scientific education using mixed qualitative and quantitative data, showed positive results on student motivation [16]. In a study conducted by Dithmer et al. in 2016, proposed a prototype experiment to help heart patients undergo telerehabilitation using the Teledialog Method to encourage patients. Also, there is research on improving learning with the gamification method associated with social networking [17]. The implementation of gamification methods and social media is to increase student participation in learning for emergency medicine in the world of health [18]. The effect of the gamification learning model also shows a positive attitude towards groups of students due to learning [19].

Today there are many approaches through media games and also social media with several cases in the world of medicine and health as a means of education. Also, the use of social media applications and games today contributes to the use of mobile as a healthy lifestyle media [8]. The use of social media and games as a means of learning for motivating students in learning mathematics [7]. The use of social media in education on mobile applications has also proven able to reduce communication difficulties in college. The case study by Denizalp, et al., involved 30 teachers and 20 students for 12 weeks focusing on communication between students and teachers through social media. [20].

The study of Yen et al. in 2018 that the effects of games change a person's behavior in many ways, including education and activity [21]. Also, gamification is introduced to motivate someone [22][23]. In this research, it is assumed that the gamification method was quickly driving someone because the game element makes people feel given a challenge. The use of a prototype to motivate cardiac patients consists of several challenges of patient activity based on several elements of gamification; the results show that the level of each game builds motivation for heart patients [14]. Also, in a study explaining the use of gamification methods and elements of social media proved able to make a positive relationship between educational practices and enthusiasm and communication [24]. We also include references from the literature review that explain gamification refers to information systems designed to provide the same experience and motivation as games, and consequently, seek to influence user behavior [22]. Even a study that uses virtual games in recognizing objects around can have a good impact on early childhood [25]. In some studies, elements in games have been proven to have a positive impact on users. In the study of Wang et al. in 2017 explained that the influence of badge and game time plays an essential role in motivating students in mathematics in the form of interactive games [26].

A prototype application cannot be launched to the user without testing, both the system and the benefits of knowing interactive experiences between games and users [27]. In the study of Santoso et al. in 2016 explained the use of measuring tools developed from the user experience questionnaire (UEQ) to measure the experience of users of learning applications. [28]. The user experience questionnaire (UEQ) explains the identification of 6 crucial issues with user experience parameters in testing an app to find the application usability or prototype design that will be used to see if the model is useful in solving problems [29]. Also, research on the identification of expressions based on facial mimics as well as participant opinion, which is helpful to know the user's response directly and also as a benchmark to find out the lack of applications based on user opinion [16] [30].

As mentioned earlier, education about tuberculosis for tuberculosis patients through the DOTS program is vital as a preventative measure. However, mobile tuberculosis education is widely interactive and fun. It can support more interesting tuberculosis education programs. Although in the previous literature explained that there had been many uses of social media and gamification methods in the world of education and health, there has been no research analyzing both approaches for tuberculosis education. In this case, this research aims to find out the effectiveness of tuberculosis education widely using the social-gamification method.

III. Research Questions

Tuberculosis education using the methods implemented in the DOTS program prioritizes the delivery of information on tuberculosis to patients and their families. The doctor or health worker educates by conveying orally and persuasively to the patient, and the family or the person closest to the patient. From previous research studies in the last section, we found several things that can be used as a reference in this study. Our research questions refer to active learning methods through social media and gamification that can be done by patients and users widely. There are some of the objectives of the research question:

- How to model social-gamification for tuberculosis education based on mobile user interaction?
- Can social-gamification content help convey information about tuberculosis education?
- Does the tuberculosis education social-gamification mobile application have an impact on user reactions such as motivating users in educating tuberculosis?

IV. Method and Material

This research is a type of development research by proposing a prototype of mobile application design for tuberculosis education using gamification that adopts elements of social media or social-gamification characteristics. The flow of this research stage is shown in Fig. 1.

From Fig. 1, there are three main stages of research. The first stage is the prototype pre-design stage; at this stage, we begin by direct observation of user needs in clinics (community Health centers/*Puskesmas*) and hospitals that treat patients with tuberculosis diagnoses. At this stage, we also identified prototype needs based on the principal elements of gamification and social media elements. In the next step, determine the Tuberculosis information that is needed part of the content of the prototype. The next stage is to build a prototype and test it. At this stage, the design is made based on the needs that have been identified at the pre-design phase. At this stage, participants are filling the form of testing based on the prototype to see the user's reaction to the application; several game scenarios are prepared at this stage [31]. The last step is the evaluation stage; we are giving out a questionnaire (UEQ) [10], opinion form, and observation of correspondent expressions. The results of this identification are then evaluated and analyzed.

A. *Studi Population*

To make the application run properly requires participant testing. This participant consists of participant par-prototype and post-prototype. We involve doctors, health workers, and voluntary participants. Total N = 15 for pre-prototype particles. As for the post-prototype, we involved 48 participants, consisting of participants with tuberculosis and general participants. These participants were distributed in 2 clinics (*Puskesmas*) and two hospitals in Yogyakarta and Central Java. Tuberculosis participants were 23 participants (N = 23), and this general participant included the families of tuberculosis patients or those who knew tuberculosis patients were 25 people (N=25). In addition to participants in clinics (*Puskemas*) and hospitals, we also include participants who come from the tuberculosis surveillance community on social media contacted online.

B. *Data Collection*

Data collection is part of the observation phase is divided into two parts: literature study and field study. In literature studies, we look for sources from a variety of previous studies related to this research. Among recommended journals and also studies through books about tuberculosis. Whereas in the field study, we observed directly with short interviews in clinics and hospitals, of tuberculosis patients, patients' families, health workers, and doctors. At this stage, we also take the primary supporting data as the content of the content that will be created.

In Fig. 2 shows one of the supporting data Tuberculosis information released by the health department. Tuberculosis information brochures are given to tuberculosis patients or as public information material. This brochure data and health information are then validated by the doctor and become the prototype information content. The content element is the most important because it relates to the purpose of research, namely tuberculosis education.

C. *Determine Element*

In developing this prototype, we selected several essential elements as design input. The determination of the specification of gamification design requirements follows the flow according to the needs analysis, which also includes social media elements. And identify the first game designs used and user requirements for the mobile application to be created [14] [30]. The two principals are as follows:

- Gamification elements include challenge elements, levels, badges, awards, and points, and leaderboard. The use of these elements is intended to add motivation to each game. Besides, each level is given a duration of play so that users are encouraged to continue playing.
- Social media elements, for this element, we include aspects like and share. Share in the prototype that was built; there are two objectives. One to share points in quizzes and tasks, and to share ranking status.

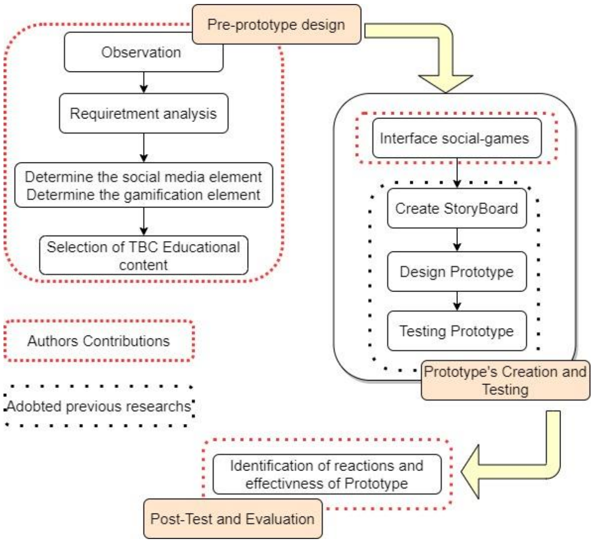

Pre-prototype design
Observation
Requiretment analysis
Determine the social media element
Determine the gamification element
Selection of TBC Educational content
Interface social-games
Create StoryBoard
Design Prototype
Testing Prototype
Prototype's Creation and Testing
Authors Contributions
Adobted previous researchs
Identification of reactions and effectivness of Prototype
Post-Test and Evaluation

Fig. 1. Research Design.

Fig. 2. Tuberculosis Brochures.

D. *Requirement Analysis*

To facilitate the needs of the system, we conducted semi-structured interviews before designing the prototype. And provide a form containing input design application features that will be made to the participant (N = 15). The purpose of this form is to determine the user's needs for application features. Participants from the doctors, health workers, and random patients selected can provide input to the system that will be created directly through this form.

TABLE. I. FEATURE AND DESCRIPTION

Feature	**Description**
Announcements	Opening the games menu to the user, and the first hint is game games.
Challenges	In the form of a quiz with tuberculosis information content.
Quiz	Other users can share quiz completion/challenge menus.
Content	The challenge is in the form of a quiz with tuberculosis information content
Share	Players can share the results of the game (achievement) their progress on social media
Profile	Players can change profile characters to avatars
Task	Completion menu about tasks consisting of activities during treatment

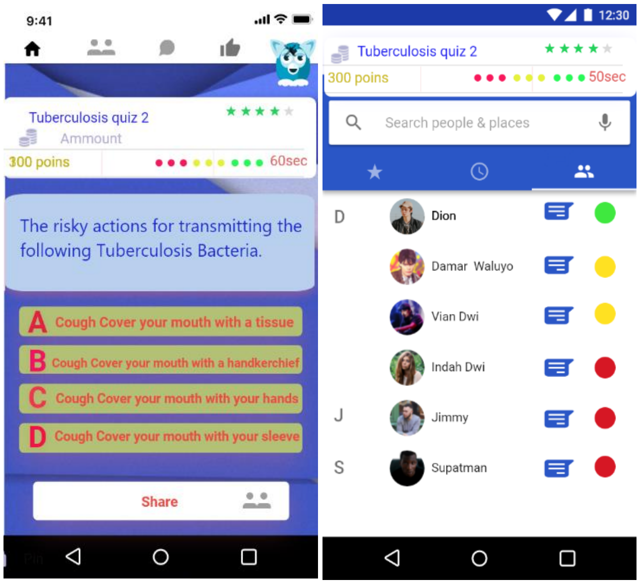
Table I is the result of the analysis of features needed based on the form that was previously randomly distributed to patients and health workers—some of the main elements required as input with details in the table. The aim of the design prototype was following the objectives, as well as knowing the needs of users with the ultimate goal of delivering tuberculosis education interactively.

V. RESULT AND DISCUSSION

To compare the performance of delivering tuberculosis educational content in the design of social-gamification applications, we prepared a game Hypothesis Scenario. One of the first scenery instruments for participant games is required to answer quizzes that are limited by backward. Each game will have a "share" help option where the player can use the help with a record of points that will be obtained only 50%, and 50% is given by friends who help. Share can also be done to show achievements in the ranking board. Badge assignment and badge redemption are given if the player has completed a game with points that meet the required amount. The prototype of the social game mobile application can be seen in Fig. 3(a & b) and Fig. 4(a & b).

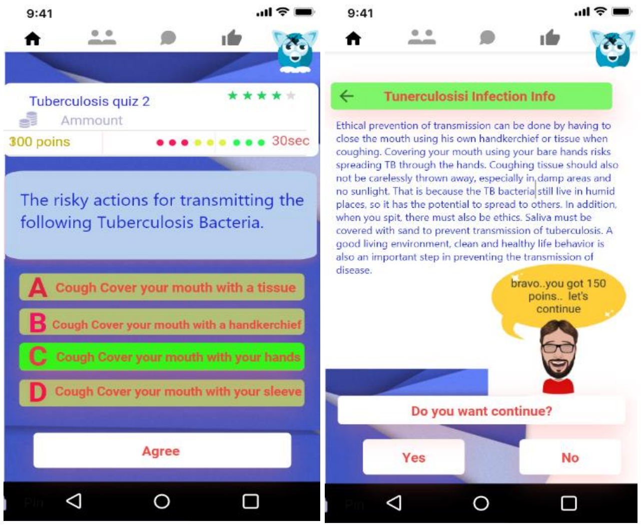
In Fig. 3(a) is one of the initial levels of games that contain quizzes with 300 points rewards. In each game, both the quiz and the player's task will be given a share option. While Fig. 3(b) is a screen that shows a list of friends' contacts. Players can choose one friend to help get useful answers.

Fig. 4(a) shows answers from a friend, in the picture, shows the time the game continues to run backward, the goal is to provide motivation. While Fig. 4(b) is a point achievement, it can be seen that the points earned are divided into 150 points. Besides that, in each final answer, there is information about the response that provides education following the target of tuberculosis education.


The risky actions for transmitting the following Tuberculosis Bacteria.

A Cough Cover your mouth with a tissue
B Cough Cover your mouth with a handkerchief
C Cough Cover your mouth with your hands
D Cough Cover your mouth with your sleeve

Fig. 3. Interface Application (3a & 3b).


The risky actions for transmitting the following Tuberculosis Bacteria.

A Cough Cover your mouth with a tissue
B Cough Cover your mouth with a handkerchief
C Cough Cover your mouth with your hands
D Cough Cover your mouth with your sleeve

Tunerculosisi Infection Info

Ethical prevention of transmission can be done by having to close the mouth using his own handkerchief or tissue when coughing. Covering your mouth using your bare hands risks spreading TB through the hands. Coughing tissue should also not be carelessly thrown away, especially in damp areas and no sunlight. That is because the TB bacteria still live in humid places, so it has the potential to spread to others. In addition, when you spit, there must also be ethics. Saliva must be covered with sand to prevent transmission of tuberculosis. A good living environment, clean and healthy life behavior is also an important step in preventing the transmission of disease.

Do you want continue?

Fig. 4. Interface Application (4a & 4d).

A. *User Experiment*

In addition to content that contains information about tuberculosis to get an interactive education, we believe that through games and social participants will be more motivated. The use of elements in games such as levels, achievement points, and badges) triggers enthusiastic participants in finding information about tuberculosis. To see an acceptable prototype application, we propose a user experience testing scheme. Testing is done on voluntary participants N = 48. During the test, we observed user behavior, especially facial expressions. In this test, we provide two simple scenarios for players.

- First Scenario, players perform quiz challenges and do tasks in the game without the help of "sharing".
- The second player's scenario is allowed to take share options to friends on the friends' list.

The results obtained, many participants who could not answer the quiz and complete the task without the help of shares, especially the general participants. While in the second scenario, many participants take the risk of answering quiz and task challenges by asking in the share options even though the answers are not always correct. From the observations of many participants with tuberculosis conditions who are not confused with the material content. Facial expressions tend to be confident in being able to answer, in contrast to most general participants. In Fig. 4, the testing process in several clinics and hospitals has a DOTS corner.

Fig. 5 shows some participants, consisting of tuberculosis patients and general participants, was testing the application. We do prototype testing to the participant by prioritizing the leading participant, namely the hypnotic tuberculosis condition. Also, participants with high-risk situations were exposed to tuberculosis bacteria in the second category, even though we included them in the general classification. The number of voluntary participants is 48 (N = 48). From the results of the prototype test using a user experiential questioner (UEQ) [10][28], we identified participant experience with six scales containing 26 items. From the application material that has been tested by the participant, we use Confidence intervals (p = 0.05) per scale to determine the measurement of the accuracy of each estimated average range. The smaller the confidence interval will be making higher accuracy of the estimate. The following are the results of user experience testing in Table II.

In Table II is the result of the questionnaire's answer to the testing of participant experience, with statistical calculations for the average scale and the average item. There are six rating scales in the user experience questioner, and our identification shows that the stimulation rating scale has the highest results with an average score of 1.578. This stimulation rating scale has a composition of items such as valuable/inferior, boring/outgoing, not interesting/interesting, motivating/decreasing motivation. Overall the results of the user experience questioner, if viewed graphically, can be seen in the graft benchmarks in Fig. 6.

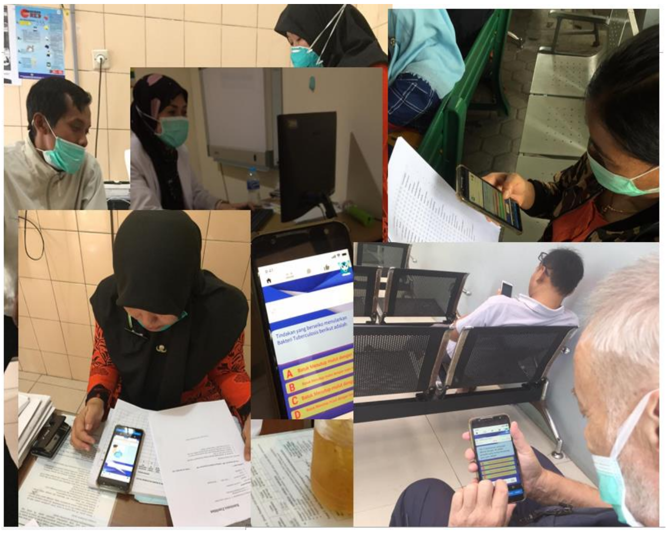


Fig. 5. Testing Application to Participants.

TABLE. II. CONFIDENCE INTERVAL RESULT

Scale	Mean	SD	N	Confidence	Confidence interval	
Attractiveness	1.483	0.862	48	0.244	1.239	1.727
Perspicuity	1.224	1.092	48	0.309	0.915	1.533
Efficiency	1.125	1.038	48	0.294	0.831	1.419
Dependability	1.083	0.876	48	0.248	0.836	1.331
Stimulation	1.578	0.959	48	0.271	1.307	1.850
Novelty	0.964	0.904	48	0.256	0.708	1.219

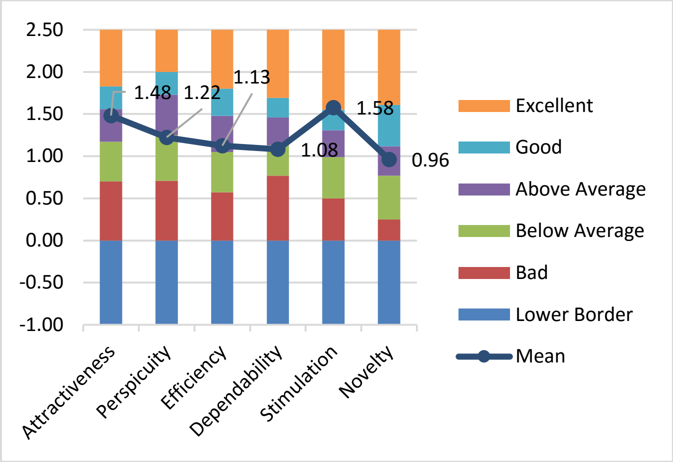


Fig. 6. Benchmark Graft All Participants.

Fig. 6 shows a comparison of the overall six rating scales. Of all items, our findings focus on identifying motivations that are compositionally stimulating items. Also on the graph can be seen the results of perspicuity with a mean score of 1.22, which is above average quality, being aware of the effectiveness of the product. Overall, the application is acceptable because most of the scores are above average, even though there is an item rating scale that is below average, namely the dependability rating scale with a mean value of 1.08. We identified that contributions to the dependability assessment scale were due to applications that sometimes did not respond as well. However, we are satisfied with the results obtained because they are under the research objectives of knowing the level of motivation and effectiveness that the application is going well.

B. Participants Opinion

In addition to interpreting the results of the user experience questionnaire statistically, to detect the prototype function is running well and as an evaluation of usability. To see the needs of the system in the future, we also provide a participant opinion form that presents insights into the experience of the participants during testing interviews and relevant quotes. To better know and define useful applications for tuberculosis education. There are two things that we describe, namely, components that are favoured by participants and elements that are not supported by participants. Due to the optional nature of the assessment carried out by 20 participants from 48 participants who are willing to provide direct evaluation and opinion. The results of identifying participant opinions are shown in Table III.

Based on Table III, we evaluate the opinions of 20 participants who are part of the voluntary participant N = 48. Some opinions about component features can be obtained; it appears that the component that participant likes is the function and usefulness of nine people, and no one dislikes it. The contribution of this component, according to participants, because the social-gamification content has educational content that can be useful, whereas the participant that is not liked is another component item with a total of seven participants. However, the second component that is not liked is the application response, with the number of six participants. Here are some participant's opinions after feeling the experience test on the application.

C2: "The content presented is exciting and easy to understand; first, I doubt the information about tuberculosis can be used as a game" (like).

C1: "Overall, this application can be used, but does not respond quickly. What I like is there is a share option, this motivates me to see what it says" (like).

C5: "hmmm, the timer at the task challenge is too fast. Only 60 seconds, I did not think, especially I do not know much about tuberculosis" (dislike).

C4: "the application is loading a long time when I choose the answer the system does not run quickly" (dislike).

C2: "I like the contents of the information; in the end, the answers are given information in detail. Even though I answered wrongly, this is very educative. I think this can make entertainment and information for everyone" (like).

C2: "I have extra-pulmonary tuberculosis, and I feel this kind of interactive information is very useful, maybe it can be a daily challenge. The data is helpful for tuberculosis sufferers who are undergoing treatment" (like).

In the participant's opinion, it can be seen that the majority of participants received the prototype positively. We identified several significant sentences, especially positive opinions. Some of the words in the participant's affirmative sentences were delivered, such as "very interesting," "motivating," "interactive," "educative," and so on. Also, we identified several opinions that tended to be negative, and this was primarily aimed at applications. Some negative sentences that we can identify, such as the use of the words "confuse" and "long loading". From this input of participant opinion, we can also find out and improve applications in future research.

TABLE. III. COMPONENTS OPINIONS FOR PARTICIPANTS

Components	Frequency	
	Like	Dislike
(C1) Games desain (in general)	3	2
(C2) Content	5	0
(C3) Function/useful	9	0
(C4) Timer to play	0	3
(C5) App respond	0	6
(C6) Etc	3	7

C. Discussion

In general, from the experience test from N = 47, participants consisting of N1 (participant with tuberculosis conditions) went well. Although there is a rating scale that has a rate below average in UEQ benchmark, the items in the Dependability rating scale, which include unpredictable/ predictable, obstructive/supportive, secure/not secure, meet expectations/does not meet expectations items, indicate that an application cannot be separated from attention in terms of security, predictions, and application interactions towards users. Besides the app, we also find indications in terms of participant population that need to be discussed. In this study, we obtained voluntary participants of N = 48, which consisted of 23 participants with a tuberculosis diagnosis of 25 general participants. The results of the answers to these two participant groups identified interesting findings to discuss. The following is a visualization of the comparison of the two-class categories seen in the comparison of scale means diagram Fig. 7.

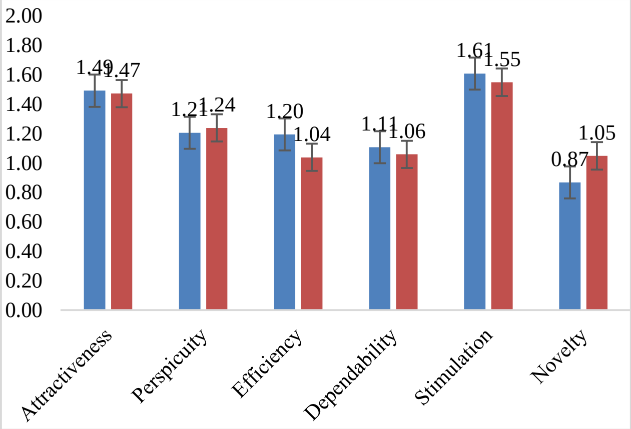


Fig. 7. Comparison of Scale Seans.

Fig. 7 shows a diagram containing a comparison of the results of participant answers. We group tuberculosis and general participant participants. There is not much difference in contrast between users with indications of Tuberculosis and typical participants. However, in some items, it appears that voluntary participants with a hint of tuberculosis are more enthusiastic compared to general participants. Of the six grading scales, four grading scales address participants with tuberculosis having higher outcomes than the general participant, namely for the rating scale of attractiveness, efficiency, dependability, and stimulation. Whereas the rating scale for perspicuity and novelty for general patients is larger. But overall, there were no significant differences in all the user experience questioner (UEQ) rating scales.

VI. Conclusion

Based on our findings in the previous section, we conclude that by adopting social media and gamification into tuberculosis education on a mobile basis, the delivery of information is more effective. The results obtained indicate that the level of user acceptance in the application is quite good. Using UEQ tools, we can identify user responses and find two essential notes, namely:

1) An assessment of the reactions of all participants. On the six UEQ ranking scales, the stimulation ranking scale has an excellent score of 1.58. Where in the stimulation assessment range, there are interesting and motivating assessment items. These findings indicate that the new and interactive education system is preferred by users and stimulates both in terms of TB education content and application features. While the attractiveness, perspective, efficiency, and novelty rating scales have an above-average rating, only the dependency rating scale has a below-average rating with a score of 1.08.

2) Based on the comparative assessment of the two participant populations, we note that there are no significant differences on the general rating scale. From the attractiveness, rating scale show that participants with tuberculosis condition have a comparison score above the general participant of 1.49 versus 1.47. While stimulation is 1.61 versus 1.55, this proves TB participants are more interested in using prototypes.

In addition to identifying statistically through the UEQ tool, we also evaluate input from participants' opinions. This opinion can be used as a reference for future studies. In future research, the use of virtual reality in daily challenges that adopt TBC treatment activities and patient activities in the DOTS program as a form of further education and assistance for TBC patients.

References

[1] "Global Tuberculosis Report 2017," 2017.

[2] G. Gebrezgabiher, G. Romha, E. Ejeta, G. Asebe, E. Zemene, and G. Ameni, "Treatment outcome of tuberculosis patients under directly observed treatment short course and factors affecting outcome in southern Ethiopia: A five-year retrospective study," PLoS One, vol. 11, no. 2, pp. 1–10, 2016.

[3] A. Probandari et al., "The path to impact of operational research on tuberculosis control policies and practices in Indonesia," Glob. Health Action, vol. 9, no. 1, 2016.

[4] I. Wayan Gede Artawan Eka Putra et al., "The implementation of early detection in tuberculosis contact investigation to improve case finding," J. Epidemiol. Glob. Health, vol. 9, no. 3, pp. 191–197, 2019.

[5] C. Maican, R. Lixandroiu, and C. Constantin, "Interactivia.ro - A study of a gamification framework using zero-cost tools," Comput. Human Behav., vol. 61, pp. 186–197, 2016.

[6] A. M. Price, K. Devis, G. LeMoine, S. Crouch, N. South, and R. Hossain, "First year nursing students use of social media within education: Results of a survey," Nurse Educ. Today, vol. 61, pp. 70–76, 2018.

[7] P. Juric, M. B. Bakaric, and M. Matetic, "Design and implementation of anonymized social network-based mobile game system for learning mathematics," Int. J. Emerg. Technol. Learn., vol. 13, no. 12, pp. 83–98, 2018.

[8] M. M. Nour, A. S. Rouf, and M. Allman-Farinelli, "Exploring young adult perspectives on the use of gamification and social media in a smartphone platform for improving vegetable intake," Appetite, vol. 120, pp. 547–556, 2018.

[9] L. Hakulinen, T. Auvinen, and A. Korhonen, "The effect of achievement badges on students' behavior: An empirical study in a university-levell computer science course," Int. J. Emerg. Technol. Learn., vol. 10, no. 1, pp. 18–29, 2015.

[10] M. Schrepp and A. Hinderks, "Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience," vol. 8517, no. June, 2014.

[11] A. Surya et al., "Quality tuberculosis care in Indonesia: Using patient pathway analysis to optimize public-private collaboration," J. Infect. Dis., vol. 216, no. suppl_7, pp. S724–S732, Nov. 2017.

[12] C. K. Lam, K. M. G. Pilote, A. Haque, J. Burzynski, C. Chuck, and M. Macaraig, "Using video technology to increase treatment completion for patients with latent tuberculosis infection on 3-month isoniazid and rifapentine: An implementation study," J. Med. Internet Res., vol. 20, no. 11, 2018.

[13] I. V., E. Nikulchev, A. A., and A. Y., "Study of Gamification Effectiveness in Online e-Learning Systems," Int. J. Adv. Comput. Sci. Appl., vol. 6, no. 2, pp. 71–77, 2015.

[14] M. Dithmer et al., "'The heart game': using gamification as part of a telerehabilitation program for heart patients," Games Health J., vol. 5, no. 1, pp. 27–33, 2016.

[15] R. Garett and S. D. Young, "Health care gamification: A study of game mechanics and elements," Technol. Knowl. Learn., pp. 1–13, 2018.

[16] C. Hursen and C. Bas, "Use of gamification applications in science education," Int. J. Emerg. Technol. Learn., vol. 14, no.1, pp. 4–23, 2019.

[17] A. M. Toda, R. M. C. do Carmo, A. P. da Silva, I. I. Bittencourt, and S. Isotani, "An approach for planning and deploying gamification concepts with social networks within educational contexts," Int. J. Inf. Manage., vol. 46, no. May, pp. 294–303, 2019.

[18] T. de A. G. Grangeia, B. de Jorge, D. Cecílio-Fernandes, R. A. Tio, and M. A. de Carvalho-Filho, "Learn+fun! Social media and gamification sum up to foster a community of practice during an emergency medicine rotation," Heal. Prof. Educ., pp. 1–15, 2018.

[19] I. Varannai, P. Sasvari, and A. Urbanovics, "The Use of Gamification in Higher Education: An Empirical Study," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 10, pp. 1–6, 2017.

[20] F. O. Hasan Denizalp, "Determination of student opinions on usage of social media and mobile tools in student-teacher, student-student communication," Int. J. Emerg. Technol. Learn., vol. 14, no. 22, pp. 19–28, 2019.

[21] B. T. H. Yen, C. Mulley, and M. Burke, "Gamification in transport interventions: Another way to improve travel behavioural change," Cities, vol. 85, no. January, pp. 140–149, 2019.

[22] J. Koivisto and J. Hamari, "The rise of motivational information systems: A review of gamification research," Int. J. Inf. Manage., vol. 45, no. July 2018, pp. 191–210, 2019.

[23] C. N. De Freitas, "Lean-based enterprise gamification: Realization of effective gamification in an enterprise context," pp. 1–48, 2015.

[24] G. Aydin, "Adoption of Gamified Systems," Int. J. Online Mark., vol. 5, no. 3, pp. 18–37, 2015.

[25] N. Kristianti, S. Niwayan Purnawati, and Suyoto, "Virtual education with puzzle games for early childhood: A study of Indonesia," Int. J. Eng. Pedagog., vol. 8, no. 2, pp. 14–22, 2018.

[26] F. Wang, Y. Wang, and X. Hu1, "Gamification teaching reform for higher vocational education in china: A case study on layout and management of distribution center," Int. J. Emerg. Technol. Learn., vol. 12, no. 9, pp. 130–144, 2017.

[27] M. Gonzalez-Salazar, H. Mitre-Hernandez, and C. Lara-Alvarez, "Method for Game Development Driven by User-eXperience: a Study of Rework, Productivity and Complexity of Use," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 2, pp. 394–402, 2017.

[28] H. B. Santoso, M. Schrepp, R. Yugo Kartono Isal, A. Y. Utomo, and B. Priyogi, "Measuring user experience of the student-centered E-learning environment," J. Educ. Online, vol. 13, no. 1, pp. 1–79, 2016.

[29] C. Maitland et al., "Measuring the capacity of active video games for social interaction: The Social Interaction Potential Assessment tool," Comput. Human Behav., vol. 87, pp. 308–316, 2018.

[30] C.-C. (Brian) Chen, "Gamify online courses with tools Built into your learning management system (LMS) to enhance self-determined and active learning," Online Learn., vol. 22, no. 3, pp. 41–54, 2018.

[31] L. De-Marcos, A. Domínguez, J. Saenz-De-Navarrete, and C. Pagés, "An empirical study comparing gamification and social networking on e-learning," Comput. Educ., vol. 75, pp. 82–91, 2014.

New Approach for the Detection of Family of Geometric Shapes in the Islamic Geometric Patterns

Ait Lahcen Yassine[1], Jali Abdelaziz[2], El Oirrak Ahmed[3], Abdelmalek. Thalal[4]
Youssef. Aboufadil[5], M. A. Elidrissi R[6]
Laboratory of Material Sciences, The University of Cadi Ayyad, Faculty of Sciences Semlalia, Marrakech, Morocco[1, 2, 4, 5, 6]
Laboratory of engineering information system, The University of Cadi Ayyad[3]
Faculty of Sciences Semlalia, Marrakech, Morocco[3]

***Abstract*—This article proposes a new approach to detect the family of geometric shapes in Islamic geometric patterns. This type of geometric pattern which is constructed by tracing the grids with the respect of precise measurement criteria and the concept of symmetry of a method which is called 'Hasba'. This geometric pattern generally found in the tiles which cover the floors or walls of many buildings around the Islamic world such as mosques. this article describe a new method which is based on the calculation of the Euclidean distance between the different geometric shapes which constitute the geometric Islamic pattern, in order to detect similar regions in this type of geometric pattern encountered in Islamic art.**

***Keywords*—Family geometric; shapes; Euclidean distance; 'Hasba'; geometric art; Islamic patterns**

I. Introduction

Since centuries, the Islamic world has had great decorative traditions. The Islamic geometric patterns [1] were very widespread throughout different countries in Africa, Asia and Europe. This article focus on the ornamental Arabesque (Thalal & al. [1-2]). More precisely, in this paper will be interested on the geometric patterns called "Tastyr" (Fig. 1). This kind of Islamic patterns are built through a method called 'Hasba' based on the respect of some precise criteria of measurement.

In Morocco, the method to build the geometric patterns adopted by craftsmen is called 'Hasba' [2], based on rigorous geometric rules such as the distance between two close neighboring geometric shapes must be fixed. The concept of 'symmetry', also based on a specific measure, is rather adapted in the carving and painting of wood, metal and plaster.

The main goal of the proposed approach consists of detecting family of similar regions through the analysis of pictures constructed based on the method 'Hasba', which containing geometric shapes. This method called "Hasba" (measure) is widely adopted by the Moroccan craftsmen ("Maâlam") especially who's working on wood material and handed over to their disciples.

This new approach to extract similar regions in Islamic geometric patterns is based on the detection first of the outline [3] of geometric shapes. In a second step, calculating the Euclidean distance [4] between them, in order to help artisans to make a decision very quickly on the geometric pattern, does it respect the rules of the 'Hasba' method ? as for example the distance between two neighboring forms is constant and the concept of symmetry is also respected.

After that the "Maâlam" can adapted the geometric pattern on wood or in other material (plaster, metal, marble...). To do so, the method of simple blob detector [5-11] is adapted. This method of having the different regions in a picture has in the input an image with a grid level [12]. The output is a set of regions.

The paper is organized as follows: In Section 2, related work. In Section 3, describe the method simple blob detector. In Section 4, shows the procedure to follow to apply the proposed approach. Section 5 contains the conclusion and the future works.

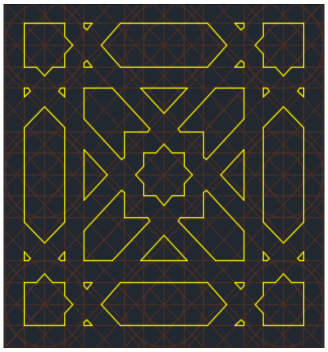


Fig. 1. Geometric Islamic Pattern.

II. Related Work

We have find articles that talk about the Hasba method, which use some software to create the models respecting 'Hasba' method. But none articles talk about how to solve the difficulty of the validation of a geometric pattern adapted by craftsmen by respected the concept of symmetry and the new Islamic geometric pattern witch take on consideration the rules of 'Hasba' method. This article talk about a new approach that based on given the steps to follow in order to give the "Maâlam" the possibility to take the decision of the geometric pattern is valid to adapt in a wood or not.

III. Simple Blob Detector

To determine the different geometric shapes, this method [13-17] uses a binary picture that allows extracting the different regions that make up the picture represent the geometric pattern. The method can follow two ways of extracting the different regions either by using 4-connectivity as shown in Fig. 2 or 8-connectivity as shown in Fig. 3.

A. Connectivity-4

This way of traversing a picture to find the different contours of the geometric shapes in the picture makes it possible to determine the corresponding neighboring pixels.

The connectivity-4 is based on the approach by flood fill that takes a binary picture in the input as shown in Fig. 4. Which represent a binary picture [18] where the value 'true' represents '1' and the value 'false' represents 0. It gives the different shapes in the output in the form of a matrix of labels as shown in Fig. 5.

In Fig. 5, the different geometric shapes in the picture represented with numeric values 0, 1 and 2. This means that the picture represent the geometric pattern, contains three regions.

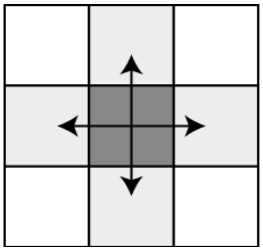


Fig. 2. Connectivity-4.

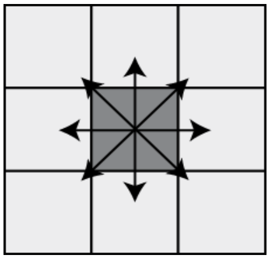
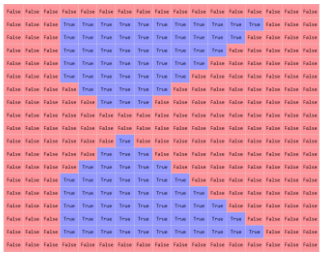
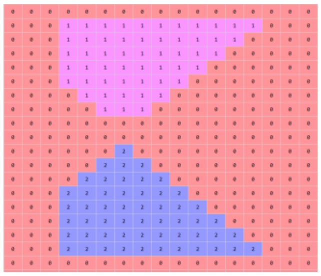


Fig. 3. Connectivity-8.



False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
False	False	False	True	True	True	True	True	True	True	True	True	True	True	False	False	False
False	False	False	True	True	True	True	True	True	True	True	True	True	False	False	False	False
False	False	False	True	True	True	True	True	True	True	True	True	False	False	False	False	False
False	False	False	True	True	True	True	True	True	True	True	False	False	False	False	False	False
False	False	False	True	True	True	True	True	True	True	False	False	False	False	False	False	False
False	False	False	False	True	True	True	True	True	False	False	False	False	False	False	False	False
False	False	False	False	False	True	True	True	False	False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
False	False	False	False	False	False	True	False	False	False	False	False	False	False	False	False	False
False	False	False	False	False	True	True	True	False	False	False	False	False	False	False	False	False
False	False	False	False	True	True	True	True	True	False	False	False	False	False	False	False	False
False	False	False	True	True	True	True	True	True	True	False	False	False	False	False	False	False
False	False	False	True	True	True	True	True	True	True	True	False	False	False	False	False	False
False	False	False	True	True	True	True	True	True	True	True	True	False	False	False	False	False
False	False	False	True	True	True	True	True	True	True	True	True	True	False	False	False	False
False	False	False	True	True	True	True	True	True	True	True	True	True	True	False	False	False
False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False

Fig. 4. Binary Picture (True (1) or False (0).



0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0
0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0
0	0	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0
0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0
0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	2	2	2	0	0	0	0	0	0	0	0	0
0	0	0	0	2	2	2	2	2	0	0	0	0	0	0	0	0
0	0	0	2	2	2	2	2	2	2	0	0	0	0	0	0	0
0	0	0	2	2	2	2	2	2	2	2	0	0	0	0	0	0
0	0	0	2	2	2	2	2	2	2	2	2	0	0	0	0	0
0	0	0	2	2	2	2	2	2	2	2	2	2	0	0	0	0
0	0	0	2	2	2	2	2	2	2	2	2	2	2	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 5. Numeric Labels of each Regions (0, 1 and 2).

1) Algorithm of the approach by flood fill: The approach by flood fill based on two matrices, the first one representing the binary picture. It can be called A. The second Matrix represents the different regions founded in the picture, called B. The main algorithm Fig. 6.

B. Connectivity-8

The approach by double course that takes the binary picture in the input and it gives as output the different regions in the form of a matrix of labels as shown in Fig. 5. This second principle scan order of raster picture based on the classical sense as shown in Fig. 6. In addition, the second course is in the opposite direction.

1) Algorithm of the approach double course: The approach double course takes two matrices, the first one represents the binary picture previously named A. The second one represents the regions already named B. The main algorithm used in this approach to extract the geometric shapes as shown in Fig. 7.

C. Results

The identification of the regions, using the Simple Blob Detector method, takes in the input (see Fig. 8) a picture in grayscale and gives in the output the two shapes in the original picture (see Fig. 9 and Fig. 10).

The different shapes with basic proprieties(x_1^i, y_1^i), which represent the coordinate's point of the top left corner, (x_2^i, y_2^i) represent the coordinates of the point in the bottom right corner, for the Shape^i which i represent the number of the first shape as shown in Fig. 9.

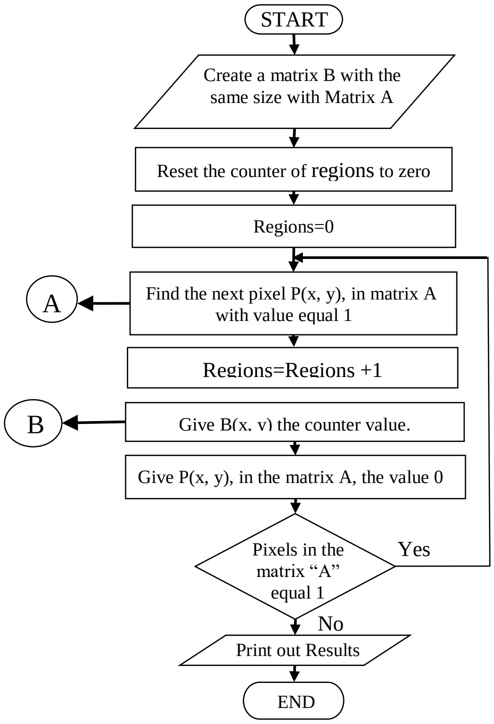

START
Create a matrix B with the same size with Matrix A
Reset the counter of regions to zero
Regions=0
A
Find the next pixel P(x, y), in matrix A with value equal 1
Regions=Regions +1
B
Give B(x, y) the counter value.
Give P(x, y), in the matrix A, the value 0
Pixels in the matrix "A" equal 1
Yes
No
Print out Results
END

Fig. 6. Main Flow Chart of the Algorithm used in this Approach by Flood Fill with Connectivity-4.

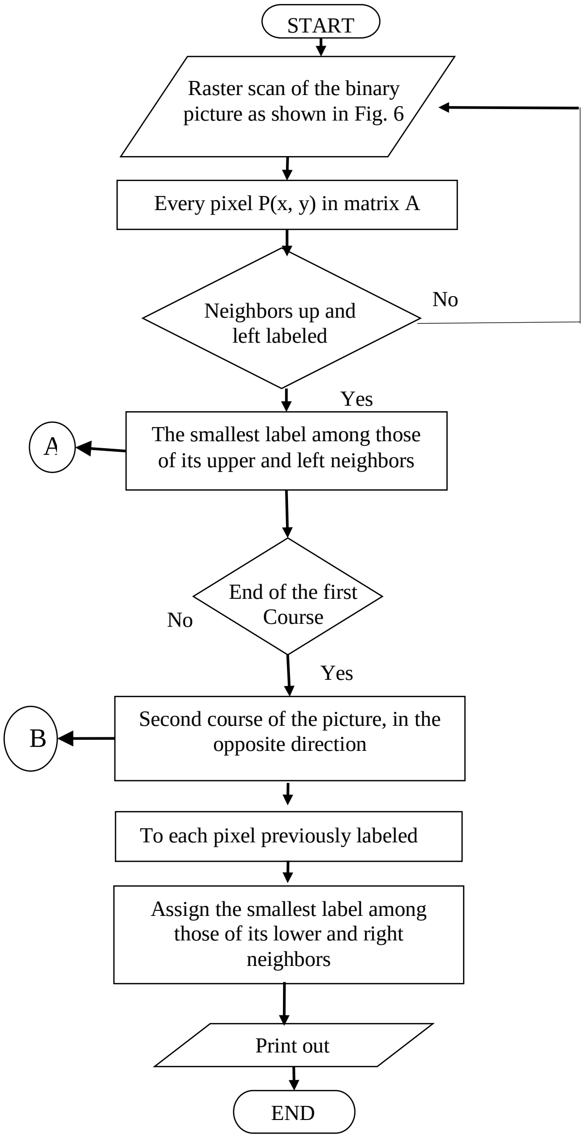

START
Raster scan of the binary picture as shown in Fig. 6
Every pixel P(x, y) in matrix A
Neighbors up and left labeled
No
Yes
The smallest label among those of its upper and left neighbors
A
End of the first Course
No
Yes
Second course of the picture, in the opposite direction
B
To each pixel previously labeled
Assign the smallest label among those of its lower and right neighbors
Print out
END

Fig. 7. Main Flow Chart of the Algorithm used in this Approach by Flood Fill with Connectivity-8.

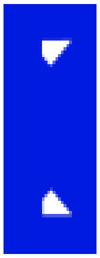


Fig. 8. Input Picture.

(x_1^j, y_1^j) represent the coordinates of the point in the top left corner, (x_2^j, y_2^j) represent the coordinates of the point in the bottom right for the Shape^j which j represent the number of the second shape as shown in Fig. 10.

In Table I, presented the x and y coordinates of the top left corner and the bottom right one for the two shapes, as shown in Fig. 11.

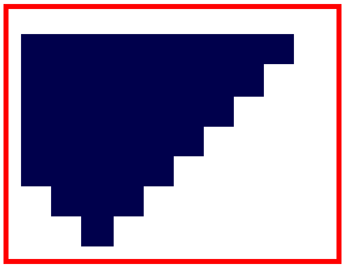


Fig. 9. Shape 1 Detected from Original Image.

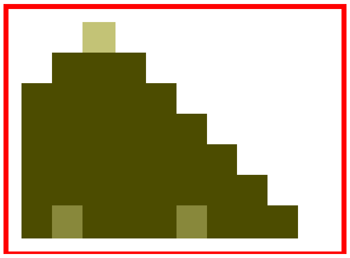


Fig. 10. Shape 2 Detected from Original Image.

TABLE. I. COORDINATES X AND Y FOR SHAPE

Coordinates	**Shape**	
	shape^i	***shape***^{j}
Coordinates x_1	2	11
Coordinates y_1	8	18
Coordinates x_2	9	19
Coordinates y_2	19	19

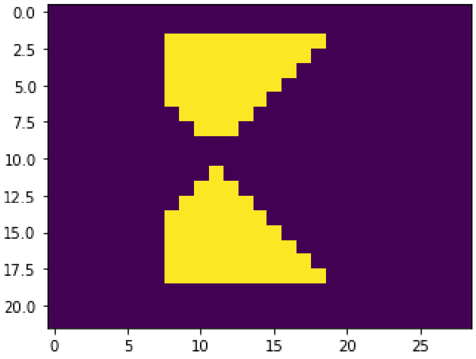


Fig. 11. Basic Proprieties ($\boldsymbol{x_1, y_1, x_2, y_2}$) for each Shapes.

IV. PROPOSED APPROACH

In this work, the Simple Blob Detector method is used with connectivity-8 to extract the different regions [8-9] in a Hasba picture Fig. 12. In order to solve the validation problem in this kind of Islamic geometric pattern, which is the distance between two neighboring geometric shapes must be the same.

The proposed approach is focus on the first part of the problem, which is how to extract similar shapes composing a family, as shown in Fig. 13.

Now, the experiment test of the approach based on detecting the different geometric shapes in 'Hasba' picture, and compute the Euclidian distance between every couple of shapes Fig. 14 and Fig. 15.

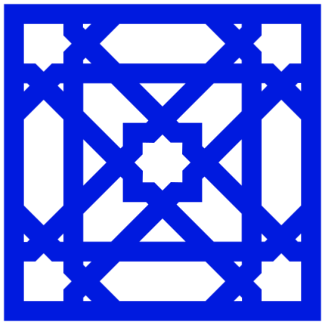


Fig. 12. Input Image Hasba.

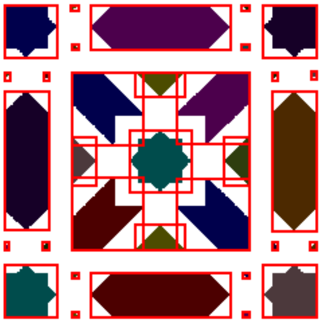


Fig. 13. Geometric Shapes Detected.

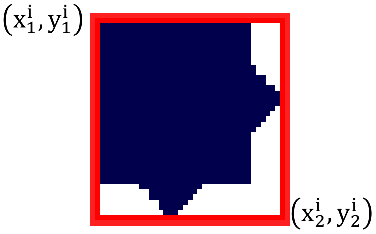


Fig. 14. Geometric Shape Characterized by (x_1, y_1) and (x_2, y_2).

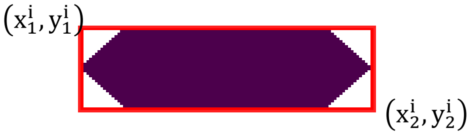


Fig. 15. Geometric Shape Characterized by (x1, y1) and (x2, y2).

A. Similarity between Regions

In Table II, represent the coordinates (x_1, y_1) and (x_2, y_2) of the different points for the every shape^i detected in the 'Hasba' picture with the Simple blob detector method.

After having detected the different coordinates x and y for each shape in the Islamic geometric pattern. The next step is to compute the Euclidian distance, using Eq. (1), between the shape^i, with coordinates $[(x_1^i, y_1^i), (x_2^i, y_2^i)]$ and the shape^j with coordinates $[(x_1^j, y_1^j), (x_2^j, y_2^j)]$ where (i, j) ϵ [1, n] and n represent the number of shapes in the Islamic geometric pattern.

$$D^1_{i,j} = \sqrt{\left(x_1^i - x_1^j\right)^2 + \left(y_1^i - y_1^j\right)^2} \qquad (1)$$

In addition, the second distance using Eq. (2).

$$D^2_{i,j} = \sqrt{\left(x_2^i - x_2^j\right)^2 + \left(y_2^i - y_2^j\right)^2} \qquad (2)$$

1) Proposed Distances used to compare the different shapes

If ${}_{i,j}^{1}D \cong {}_{i,j}^{2}D$, that means the two shapes are not necessarily similar.

After some tests, founded that there was no equality between the two distances $[D^1_{i,j}\ , D^2_{i,j}]$ and the two shapes [shape^i, shape^j] are similar. As shown in Table III.

TABLE. II. COORDINATES (X_1, Y_1) AND (X_2, Y_2)

shapes^i	**Coordinates**			
	x_1^i	y_1^i	x_2^i	y_2^i
Shape 1	27	27	89	89
Shape 2	27	105	33	114
Shape 3	27	128	79	292
Shape 4	27	305	34	314
Shape 5	27	331	89	393
Shape 6	73	106	79	114
Shape 7	73	305	79	314
Shape 8	105	387	114	393
Shape 9	106	27	115	34
Shape 10	106	73	115	79
Shape 11	106	106	191	190
Shape 12	106	181	135	239
Shape 13	106	229	190	314
Shape 14	106	341	115	347
Shape 15	128	27	291	79
Shape 16	128	341	292	393
Shape 17	174	174	246	246
Shape 18	181	106	239	135
Shape 19	181	285	239	314
Shape 20	229	106	314	190
Shape 21	230	230	314	314
Shape 22	285	181	314	239
Shape 23	305	27	314	33
Shape 24	305	72	314	79
Shape 25	305	341	314	347
Shape 26	305	386	314	393
Shape 27	331	27	393	89
Shape 28	331	331	393	393
Shape 29	341	106	347	115
Shape 30	341	128	393	292
Shape 31	341	306	347	314
Shape 32	386	106	393	115
Shape 33	387	306	393	315

TABLE. III. LIST OF THE DISTANCES $D^1_{i,j}$ AND $D^2_{i,j}$ FOR THE SHAPES

shapes	**Distances $D^1_{i,j}$ and $D^2_{i,j}$**			
	Shape^i	Shape^j	$D^1_{i,j}$	$D^2_{i,j}$
Shape 1	1	1	0	0
Shape 2	1	5	304.0	304.0
Shape 3	1	27	304.0	304.0
Shape 4	1	28	429.92	429.92
Shape 5	2	2	0	0
Shape 6	2	4	200.0	200.0
Shape 7	2	6	46.01	46.01
Shape 8	2	7	205.22	205.22
Shape 9	2	8	292.58	290.52
Shape 10	2	9	111.01	114.56
Shape 11	2	10	85.2	89.15
Shape 12	2	14	248.87	247.0
Shape 13	2	23	288.73	292.44
Shape 14	2	24	279.95	283.1
Shape 15	2	25	364.66	365.0
Shape 16	2	26	395.27	395.98
Shape 17	2	29	314.00	314.00
Shape 18	2	31	372.82	372.28
Shape 19	2	32	359.0	360.0
Shape 20	2	33	412.3	412.3
Shape 21	3	3	0	0
Shape 22	3	15	142.83	300.5
Shape 23	3	16	235.73	235.73
Shape 24	3	30	314.0	314.0
Shape 25	11	11	0	0
Shape 26	11	13	123.0	124
Shape 27	11	20	123.0	123.0
Shape 28	11	21	175.36	174.65
Shape 29	12	12	0	0
Shape 30	12	18	106.06	147.07
Shape 31	12	19	128.22	128.22
Shape 32	12	22	179.0	179.0
Shape 33	17	17	0	0

To solve this problem, the distances in Eq. (3) and Eq. (4) present in Table IV are added to solve the problem detected.

$$D^3_{i,j} = \sqrt{\left(\mathrm{x}^{\mathrm{i}}_1 - \mathrm{x}^{\mathrm{j}}_2\right)^2 + \left(\mathrm{y}^{\mathrm{i}}_1 - \mathrm{y}^{\mathrm{j}}_2\right)^2} \tag{3}$$

$$D^4_{i,j} = \sqrt{\left(\mathrm{x}^{\mathrm{i}}_2 - \mathrm{x}^{\mathrm{j}}_1\right)^2 + \left(\mathrm{y}^{\mathrm{i}}_2 - \mathrm{y}^{\mathrm{j}}_1\right)^2} \tag{4}$$

After the tests, some exceptions are founded that which the shape^i and shape^j are not completely similar although $D^1_{i,j} \cong D^2_{i,j}$or $D^3_{i,j}\cong D^4_{i,j}$.

A solution of this new problem consists on computing the distance using Eq. (5) and Eq. (6) present in Table V.

$$D^5_{i,j} = \sqrt{\left(x^i_c - x^i_1\right)^2 + \left(y^i_c - y^i_1\right)^2} \tag{5}$$

$$D^6_{i,j} = \sqrt{\left(\mathrm{x}^{\mathrm{j}}_{\mathrm{c}} - \mathrm{x}^{\mathrm{j}}_1\right)^2 + \left(\mathrm{y}^{\mathrm{j}}_{\mathrm{c}} - \mathrm{y}^{\mathrm{j}}_1\right)^2} \tag{6}$$

Where

$$x^i_c = \{x^i_1 + x^i_2\}/_2 \tag{7}$$

$$y^i_c = \{y^i_1 + y^i_2\}/_2 \tag{8}$$

$$x^j_c = \{x^j_1 + x^j_2\}/_2 \tag{9}$$

$$y^j_c = \{y^j_1 + y^j_2\}/_2 \tag{10}$$

TABLE. IV. LIST OF THE DISTANCES $D^3_{i,j}$ AND $D^4_{i,j}$ FOR THE SHAPES

shapes	**Distances $D^3_{i,j}$ and $D^4_{i,j}$**			
	Shape^i	Shape^j	$D^3_{i,j}$	$D^4_{i,j}$
Shape 1	1	1	87.68	87.68
Shape 2	1	5	371.21	371.21
Shape 3	1	27	371.21	249.81
Shape 4	1	28	517.6	342.23
Shape 5	2	2	10.81	10.81
Shape 6	2	4	209.11	191.09
Shape 7	2	6	52.77	40.79
Shape 8	2	7	215.37	195.14
Shape 9	2	8	300.85	282.33
Shape 10	2	9	113.07	113.56
Shape 11	2	10	91.76	83.72
Shape 12	2	14	257.50	238.44
Shape 13	2	23	295.89	285.57
Shape 14	2	24	288.17	275.22
Shape 15	2	25	375.41	354.27
Shape 16	2	26	406.58	384.66
Shape 17	2	29	320.15	308.10
Shape 18	2	31	382.20	362.94
Shape 19	2	32	366.13	353.09
Shape 20	2	33	421.96	402.71
Shape 21	3	3	172.04	172.04
Shape 22	3	15	268.50	269.4
Shape 23	3	16	374.76	69.29
Shape 24	3	30	401.06	309.09
Shape 25	11	11	119.50	119.50
Shape 26	11	13	224.32	93.52
Shape 27	11	20	224.32	92.19
Shape 28	11	21	294.15	55.86
Shape 29	12	12	64.84	64.84
Shape 30	12	18	140.73	140.73
Shape 31	12	19	188.09	65.05
Shape 32	12	22	215.93	160.82
Shape 33	17	17	101.82	101.82

TABLE. V. LIST OF THE DISTANCES $D^5_{i,j}$ AND $D^6_{i,j}$ FOR THE SHAPES

shapes	Distances $D^5_{i,j}$ and $D^6_{i,j}$			
	Shape^i	Shape^j	$D^5_{i,j}$	$D^6_{i,j}$
Shape 1	1	1	43.84	43.84
Shape 2	1	5	43.84	43.84
Shape 3	1	27	43.84	43.84
Shape 4	1	28	43.84	43.84
Shape 5	2	2	5.40	5.40
Shape 6	2	4	5.40	5.40
Shape 7	2	6	5.40	5.0
Shape 8	2	7	5.40	5.40
Shape 9	2	8	5.40	5.40
Shape 10	2	9	5.40	5.70
Shape 11	2	10	5.40	5.40
Shape 12	2	14	5.40	5.40
Shape 13	2	23	5.40	5.40
Shape 14	2	24	5.40	5.70
Shape 15	2	25	5.40	5.40
Shape 16	2	26	5.40	5.70
Shape 17	2	29	5.40	5.40
Shape 18	2	31	5.40	5.00
Shape 19	2	32	5.40	5.70
Shape 20	2	33	5.40	5.40
Shape 21	3	3	86.02	86.02
Shape 22	3	15	86.02	86.02
Shape 23	3	16	86.02	86.54
Shape 24	3	30	86.02	86.02
Shape 25	11	11	59.75	59.75
Shape 26	11	13	59.75	59.75
Shape 27	11	20	59.75	59.75
Shape 28	11	21	59.75	59.39
Shape 29	12	12	32.42	32.42
Shape 30	12	18	32.42	32.42
Shape 31	12	19	32.42	32.42
Shape 32	12	22	32.42	32.42
Shape 33	17	17	50.91	50.91

Now if

$$\left(D^1_{i,j} \cong D^2_{i,j}\ and\ D^5_{i,j} \cong\ D^6_{i,j}\right) or \left(D^3_{i,j} \cong D^4_{i,j}\ and\ D^5_{i,j} \cong D^6_{i,j}\right)$$

Shapei and shapej are very similar.

B. Family of Similar Shapes in the Islamic Pattern

In this paper the approach, after having extracted the similar regions [10-11], a problem related to similar regions is founded. For example, if they are three regions, F_1 with index 1, F_2 with index 6 and F_3 with index 27, the program will show that F_1 is similar to F_1, F_1 is similar to F_2, F_1 is similar to F_3, F_2 is similar to F_2, F_2 is similar to F_3 and F_3 is similar to F_3.

The objective is to eliminate the combinations F_2 is similar to F_2, F_2 is similar to F_3 and F_3 is similar to F_3.

To solve this problem, a set of points created, and testing whether the shape is already present in the set. If that it is so, then it not inserted in the set. Else, it will be insert it in the set of points. Table III shows the results with this method of computing the Euclidian distance among different regions.

In this subsection, which display the different family of geometric shapes, and the numbers of all shapes composing the family. Table VI shows results of different shapes in Hasba Pictures.

C. Units Display of the different Family Groups

In this subsection, witch display the results obtained through using this new proposed approach for the detection of the similar geometric shapes in Islamic geometric patterns or Hasba Pictures Fig. 16.

TABLE. VI. FAMILY OF EACH SHAPE

Family's	Number of shapes for each family	
	family shapes	***Number***
Family 1	1, 5, 27, 28	4
Family 2	2, 4, 6, 7, 8, 9, 10, 14, 23, 24, 25, 26, 29, 31, 32, 33	16
Family 3	3, 15, 16, 30	4
Family 4	1911, 13, 20, 21	4
Family 5	12, 18, 19, 22	4
Family 6	17	1

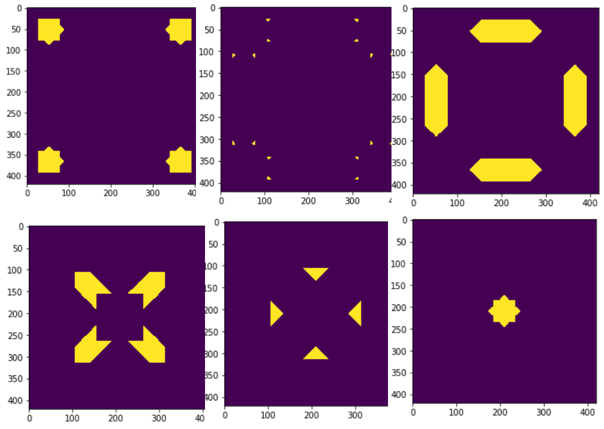


Fig. 16. Family Group of Shapes.

V. CONCLUSIONS

Many works have used the simple blob detection method to show us the different connected components in a binary picture. The problem treated in this work is not only the detection of connected components but also how to extract the similar regions in an Islamic geometric pattern. This paper presented a new approach to solve the problem: How can extract a family group of each shape in the pictures constructed with the method called 'Hasba'. In a future work, the objective is to implement new moments and apply the new detected approach to know whether a given kind of Islamic geometric patterns is valid or not.

ACKNOWLEDGMENT

The authors are grateful to Prof. Dr. Thalal Abdelmalek for pictures from Hasba presented in this paper and for fruitful discussions.

REFERENCES

[1] Thalal, A., Benatia, M. J., Jali, A., Aboufadil, Y. & Elidrissi Raghni, M. A. (2011). Symmetry Culture Sci. 22, 1-2, 103-130.

[2] Aboufadil Y, Thalal A, Elidrissi Raghni MA. Moroccan ornamental quasiperiodic patterns constructed by the multigrid method. J Appl Cryst. 2014; 47:630-641.

[3] http://www.labbookpages.co.uk/software/imgProc/blobDetection.html.

[4] https://en.wikipedia.org/wiki/Euclidean_distance.

[5] https://en.wikipedia.org/wiki/Connected-component_labeling.

[6] Azami N., Idrissi D.E., Amrane S., Harmouchi M. Computer blob detection and tracking for highly repeatable optical fiber sensor; Proceedings of the 2014 9th International Conference on Intelligent Systems: Theories and Applications (SITA-14); Rabat, Morocco. 7–8 May 2014; pp. 1–5.

[7] Philipp S., Vieira B., and Sanfourche M., "Fuzzy Segmentation of Color Images and Indexing of Fuzzy Regions," in proceedings of Conference on Colour in Graphics, Imaging, and Vision, Poitiers, pp. 507-512, 2002.

[8] Eum S., Jung H.G. Enhancing Light Blob Detection for Intelligent Headlight Control Using Lane Detection. IEEE Trans. Intell. Transp. Syst. 2013;14:1003–1011. doi: 10.1109/TITS.2012.2233736.

[9] Patro B.N. Design and implementation of novel image segmentation and BLOB detection algorithm for real-time video surveillance using DaVinci processor; Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI); Delhi, India. 24–27 September 2014; pp. 1909–1915.

[10] Wu K., Otoo E.J., Suzuki K. Optimizing two-pass connected-component labeling algorithms. Pattern Anal. Appl. 2009;12:117–135. doi: 10.1007/s10044-008-0109-y.

[11] Kiran B., Ramakrishnan K., Kumar Y., Anoop K.P. An improved connected component labeling by recursive label propagation; Proceedings of the 2011 National Conference on Communications (NCC); Bangalore, India. 28–30 January 2011; pp. 1–5.

[12] Derrode S. and Ghorbel F., "Robust and Efficient Fourier-mellin Transform Approximations for Invariant Grey-level Image Description and Reconstruction," Computer Vision and Image Understanding, vol. 83, no. 1, pp. 57-78, 2001.

[13] Chang F., jen Chen C., jen Lu C. A linear-time component-labeling algorithm using contour tracing technique. Comput. Vis. Image Underst. 2004;93:206–220. doi: 10.1016/j.cviu.2003.09.002.

[14] T. Q. CHEN, Y. LU, Color image segmentation: an innovative approach, Pattern Recognition 35 (2), 395-405, (2002).

[15] Nguyen T.B., Chung S.T. An Improved Real-Time Blob Detection for Visual Surveillance; Proceedings of the CISP '09. 2nd International Congress on Image and Signal Processing; Tianjin, China. 17–19 October 2009; pp. 1–5.

[16] Y. DENG, B. MANJUNATH, Unsupervised segmentation of colortexture regions in images and video, IEEE Trans. on Pattern Analysis and Machine Intelligence 23 (8), 800-810, (2011).

[17] Sameer A., Rangachar K., and Ramesh J., "A Survey on the Use of Pattern Recognition Methods for Abstraction, Indexing and Retrieval of Images and Video," Pattern Recognition, vol. 35, no. 4, pp. 945-965, 2002.

[18] https://fr.scribd.com/document/97422644/07-Pictures-Binaires.

NOTES ON CONTRIBUTORS

Ait Lahcen Yassine is currently a PhD student at Cadi Ayyad University, Marrakesh, Morocco, with the main field of interest being Computer Science and other fields of interest being Moroccan Art. The author's current research interest is computational method for Moroccan geometric art, digital construction of geometric patterns and modelling. The author's research activities are three Communications in the 2 last years.

Jali Abdelaziz is Ph.D. of University Higher Normal School Paris, University Paris 13, with the main field of interest being Numerical Analysis and other fields of interest being Mathematics current research interest is computational method for Moroccan geometric art, digital construction of – Computer programming infography. The author's current research interest is Moroccan Art and Computer programming. The author is Assistant Professor, Department of Mathematics, Cadi Ayyad University, Marrakesh, and Member of the Council of the Department of Mathematics, University Cadi Ayyad, and Marrakesh. The author's research activities are 7 Publications and Communications in the 10 last years, Supervision of 2 Theses, and Supervision of 12 Master's projects.

El Oirrak Ahmed joined Cadi Ayyad University, Morocco, in 1999, first as an assistant professor, and received the Doctorate and Habilitation in signal processing from the Mohammed V University, Morocco, in 2001 and University Cadi Ayyad, Morocco, in 2010 respectively. He is presently a PES professor with the Faculty of Sciences of Marrakech Semlalia. His research interests include image processing, pattern recognition and their applications. He is the author or co-author of more than 20 publications.

Thalal Abdelmalek is a Doctorat d'Etat (PhD) in Materials Sciences, P&M Curie University, Paris 6, and France. The author's main field of interest is Materials Sciences. Other fields of interest include Physics and the current research interest is Moroccan Art and Crystallography. The author is Professor, Department of Physics, Cadi Ayyad University, and Marrakech. Chairperson of the 24th European Crystallographic Meeting Conference (ECM24)–2007. President of the Moroccan Crystallographic Association (AMC), Organizer of 6 Moroccan Schools of Crystallography, Coordinator of the International Year of Crystallography Activities in Morocco: OpenLabs Morocco–2014, Chairman of the conference "Crystallography for the next generation: the legacy of IYCr" –Rabat–2015,Member of the Scientific Board of the International Basic Sciences Programme (IBSP) –UNESCO. The author's research activities include 20 Selected Publications and Communications in the 10 last years and Supervision of 10 Theses.

Aboufadil Youssef is Doctorat (PhD) in Crystallography, Cadi Ayyad University, Marrakesh, Morocco, with the main field of interest being Materials Science. Other fields of interest include Physics. The author's current research interest is Crystallography, Ceramic and Moroccan Art. The author is Assistant Professor, Department of Physics, Cadi Ayyad University, and Marrakesh. Member of the Moroccan Crystallographic Association (AMC), participant in the organization of the 24th European Crystallographic Meeting Conference (ECM24) – 2007, participant in the organization of the conference "Crystallography for the next generation: the legacy of IYCr" – Rabat – 2015 The author's research activities are 18 Publications and Communications in the 4 last years and 6 workshops in Symmetry and Geometric Art.

El Idrissi Raghni My Ahmed is Doctorat d'Etat (PhD) in Materials Sciences,Cadi Ayyad university, Marrakesh Marocco, with the main field of interest beingMaterials Science. Other fields of interest include Physics and the current research interest is Moroccan Art and Ceramic Pigment. The author is Professor, Department of Physics, Cadi Ayyad University,Marrakesh, Director of the Laboratory of Materials Sciences, Faculty of Sciences Semlalia, Cadi Ayya University, participant in the organization of the 24th European Crystallographic Meeting Conference (ECM24) – 2007. The author's research activities are 15 Publications in the 10 last years and Supervision of 4 Theses.

Place-based Uncertainty Prediction using IoT Devices for a Smart Home Environment

Dr. Amr Jadi
Department of Computer Science and Information
College of Computer Science and Engineering, University of Hail, Hail, Saudi Arabia

Abstract—**In this work, an uncertainty prediction method for the home environment is proposed using the IoT devices (sensors) for predicting uncertainties using place-based approach. A neural network (NN) based smart communication system was implemented to test the results obtained from place-based approach using the inputs from sensors linked with internet of things (IoT). In general, there are so many smart systems for home automation is available for alerting the owners using IoT, but they can communicate only after an accident happens. But it is always better to predict a hazard before it happens is very important for a safe home environment due to the presence of kids and pet animals at home in the absence of parents and guardians. Therefore, in this work, the uncertainty prediction component (UPC) using place-based approach helps to make suitable prediction decisions and plays a vital role to predict uncertain events at the smart home environment. A comparison of different classifiers like multi-layer perceptrons (MLP), Bayesian Networks (BN), Support Vector Machines (SVM), and Dynamic Time Warping (DTW) is made to understand the accuracy of the obtained results using the proposed approach. The results obtained in this method shows that place-based approach is providing far better results as compared to the global approach with respect to training and testing time as well. Almost a difference of 10 times is seen with respect to the computing times, which is a good improvement to predict uncertainties at a faster rate.**

Keywords*—*IoT; place-based approach; uncertainty prediction; MLP; SVM; BN; DTW

I. Introduction

In recent times, the internet of things (IoT) playing a pivotal role in building a modern society, infrastructure and been instrumental towards the rapid growth of smart cities [1]. The trend of people migrating to urban cities for a better lifestyle made most of the governing authorities to allow improving the growth of smart cities. Therefore, in simple, it is very much evident that major activities by humans, machinery, and administration have a great tracking of life events. Similarly, the possibility of being ignorant/least bothered, as a human tendency towards different homely activities increases due to the excessive involvement of machine-made activities in life. The possibility of these machines to get failed or giving false output response cannot be denied. Especially when it is related to kids at home in the absence of parents/guardians the possibility of risk can be more terrifying and may sometimes lead to death as well. Therefore, it is always suggested to use disruptive technologies (such as artificial intelligence, blockchain, 3D printing, IoT, etc.) carefully especially when it is applied to a home environment [2]. Yes, today's home environment is completely based on electronic gadgets/machinery and is capable of making things faster and easier for a comfortable human life. They are just one click away from making wonders with their features and quality services but a wrong click also may push you into uncertain situations as well. They need careful/skilful operators for better services and need a careful monitoring system for these device functionalities. Recently many business models are proposed to make secure and smart IoT based home environments by leading real-estate companies to enhance the comfort zone of human life.

A secured smart home was suggested by Yuan and Peng based on IoT by interfacing the web and smart phone applications for improving the user experience [3]. The authors provide a full solution to develop a smart home by using hardware design, intelligent controlling method, pervasive computing and virtual reality within their model. They also claimed to build a reduced power consumption and energy consumption model with a secured mechanism of the terminal gateway group. The health hazards using these smart homes and within the smart cities were addressed by Miori and Russo uses semantic knowledge representation using the web 3.0 [4]. The authors used specific ontology's that consider the information from distributed environments. An integration approach based on IPV6 enabled service-oriented architecture (SOA) was proposed by Jung et al. to build automation and smart cities [5]. A proof of concept implementation and performance evaluation results is produced out of this work to build an advanced control scenario for the context of smart cities. Piyare used android based smart phones to develop a monitoring and home control system with low cost and flexible mode of operation [6]. Without using a dedicated server this method proposed a novel communication protocol for monitoring and controlling the home environment with all kinds of switching functionalities using mobile devices. The biggest disadvantage of IoT based systems towards achieving security in an IoT environment. A risk analysis was carried out towards using the smart home automation systems by Jacobsson et al. by involving the leading industrial actors [7]. The results from this research indicated reducing the risks by adding standard security features to the existing IoT architectures. The IoT services provided by Vivek and Sunil are in a secured way by using the Wi-Fi ZigBee gateway for home automation [8]. In this work, this gateway includes the user interaction capabilities with an efficient way of sending and receiving the instruction from different protocols and the graphical user interface (GUI) in this work allows the users to get interacted with the settings of the ambient environment. Later, Puri and Nayyar suggested a home automation technique using PIC microcontroller (PIC 16F877A), Bluetooth (HC-05) sensors and android based technology [9]. The Bluetooth sensors used in this method are claimed to be useful for long-range and energy-efficient wireless communications. However, it might not be more appropriate using the Bluetooth for the long-range applications in the urban city environment due to the involvement of heavy noise. The role of wireless sensor networks (WSN) plays a vital role in covering the long ranges and to avoid noise issues in several applications. A secured IoT based smart home automation system was proposed by Pirbhulal et al. used the WSN for operating different home appliances [10]. In this work, for providing suitable energy-efficient data encryption the authors used a triangle-based security algorithm (TBSA), which is based on an efficient key generation mechanism. Using this method, the secured data transmission was possible and the network could cover long ranges as well. An attempt made by Saha et al. to use the advanced IoT based remote system for health monitoring, home automation and an alarm system [11]. From the proposed method a patient will get an alarm to provide the prescribed medicine in the scheduled time by using an email or SMS. An interactive dual-mode IoT based smart home automation was proposed by Hamdan et al. can monitor and control most of the home appliances remotely [12]. These appliances are interfaced with a single chip microcontroller with an in-built wireless access point which establishes the communication with the home server. This system is scalable and can add or remove the devices connected from different rooms based on the demand/ priority.

In this work, neural network (NN) based methods are used to identify and implement for any kind of uncertainty prediction with the help of IoT devices for establishing a smart communication between the owners and prediction component. The place-based uncertainty prediction models are introduced here for improving efficiency, speed, and accuracy. In the next section, detailed information of neural networks functionality as an uncertainty prediction module is presented for the smart home environment.

II. Functioning of Neural Networks

NN works on biological genetic pressures applied to pre-wire forms of the natural neural networks [13]. The first layer is formed with the input nodes followed by, hidden layers and output layers. The nodes and layers of an artificial NN (ANN) represented for a nucleus and axon of a biological NN.

A. Biological Features of Neural Networks

In a biological NN, a neuron will be surrounded with a hair-like (thin) element (i.e. dendrites), which enables the active form of a neuron. They work as input terminals for different sources with certain threshold values. These neurons will burst when the summation of all input signals reachs to maximum levels and the resultant outputs will be carried forward by the axons. These axons will be thicker in size and potentially long as compared to the dendrites, which influence remote neurons that are linked with thousands of other neurons as well. In an artificial NN, the hidden layers will be functioning like the intermediate layers, which receive the input layers and combines them based on the weights of the edges. The calculated output is emitted to the outputs by the subsequent layers, which is considered as a predicted attribute. The NN based systems can be very useful and are very good/faster risk prediction systems in most of the real-time applications such as the healthcare industry, hospitality businesses, stock market detections, etc. [14]. On the other side, the same NN systems can be used as a group for multiple parameter prediction operations such as blood pressure, sugar levels, and heartbeats in the case of the hospital environment. For example, in a home environment the room temperature, humidity, pressure level, smog, windows and door positions, water levels in a tank, electronic switch positions (ON/OFF), status of electronic accessories, automated machines, etc. must be operated within their ideal/defined status levels. These ideal/defined values will be trained to the hidden layers of the neural networks.

B. How a Neural Network Works?

The role of NN is to receive the inputs from the sensor devices that are working based on the fundamentals of internet of things (IoT) and process them along with the ideal values and trained values of the hidden layers of a NN and produce the resultant outputs at output nodes. The output values from the NN system will generate the vector outputs matrix. The conversion of these matrix values will be taking place internally to provide some of the numerical values at the output. These numerical values will be compared by the checker component with various ideal values stored in the database for predicting the uncertainty in the home environment.

To perform the internal operations within the different layers of the NN back propagation algorithm was used to determine the resultant output from the given input signals and the trained values in the hidden layers. In the back-propagation algorithm, the data will be trained to the hidden layers as per the required/expected outputs. The main reason to use this algorithm is due to its nature of being a supervised learning algorithm, which uses multi-layer perceptions for changing the weights of the hidden layers based on the adjustments needed to obtain at the output nodes [13]. This algorithm uses computed output error values to rectify the weights in the backward direction, whereas forward propagation is used to retrieve the total error in this method. Neurons will be activated at the time of forwarding propagation using a sigmoid activation function as given below:

$$f(x) = \frac{1}{(1+exp^{-input})} \quad (1)$$

The back-propagation algorithm works with four following steps:

- Using the input patterns, forward propagation is performed to calculate the error output.
- Weight values of the weight matrix will be modified based on the resultant values obtained by using Eq. 1.
- Repeat step 1.

- Process of the algorithm finishes once the output patterns are matched with the target values/patterns.

C. Uncertainty Prediction Component

The role of uncertainty prediction component (UPC) in this work ensures to obtain the information (i.e. ideal values of different parameters from the database) and to compare with the values updated at the input nodes received from sensors by converting them to input vector-matrix form as shown in Fig. 1. The hidden layer weights and random values will be defined by using the back-propagation algorithm and by using some predefined values respectively. The input vector-matrix is compared with the hidden layer weights and the resultant output vector-matrix will be formed to supply the output to the checker component.

D. Functioning of Uncertainty Prediction Component

There are different parameters considered in this work to obtain the sensor data to be analyzed and it will be in the form of analog data. The data collected from the home environment must be converted to a digital and floating form in terms of ‘0’ and ‘1’ because the neural network-based hidden layers will understand the inputs given in digital form only. The weight of the neurons will be equal to ‘0’ or ‘1’ based on the resultant processed values of inputs at nodes and the hidden layer values (i.e. trained values). The vector data set will be converted as a set of a matrix at the input layers from the obtained analog input equivalent values.

The uncertainty prediction component using NN helps to improve the speed of prediction towards any kind of uncertainty in the home environment. The problem detected by the sensing devices can be processed and identification of the risk using the proposed UPC improved the efficiency of the system. Accurate results can be obtained easily using the UPC. In the next section, the detailed architecture of the proposed method for the smart home environment is explained with suitable design components used for IoT. In simple, this method can be called as artificial intelligence (AI) based architecture uses machine learning concepts (such as multi-layer perceptrons (MLP), support vector machines (SVM), Bayesian networks (BN), and dynamic time warping (DWT), etc.) also. These techniques help in computing the inputs of different forms observed by the sensors in the absence of the human helps to predict the uncertainties.

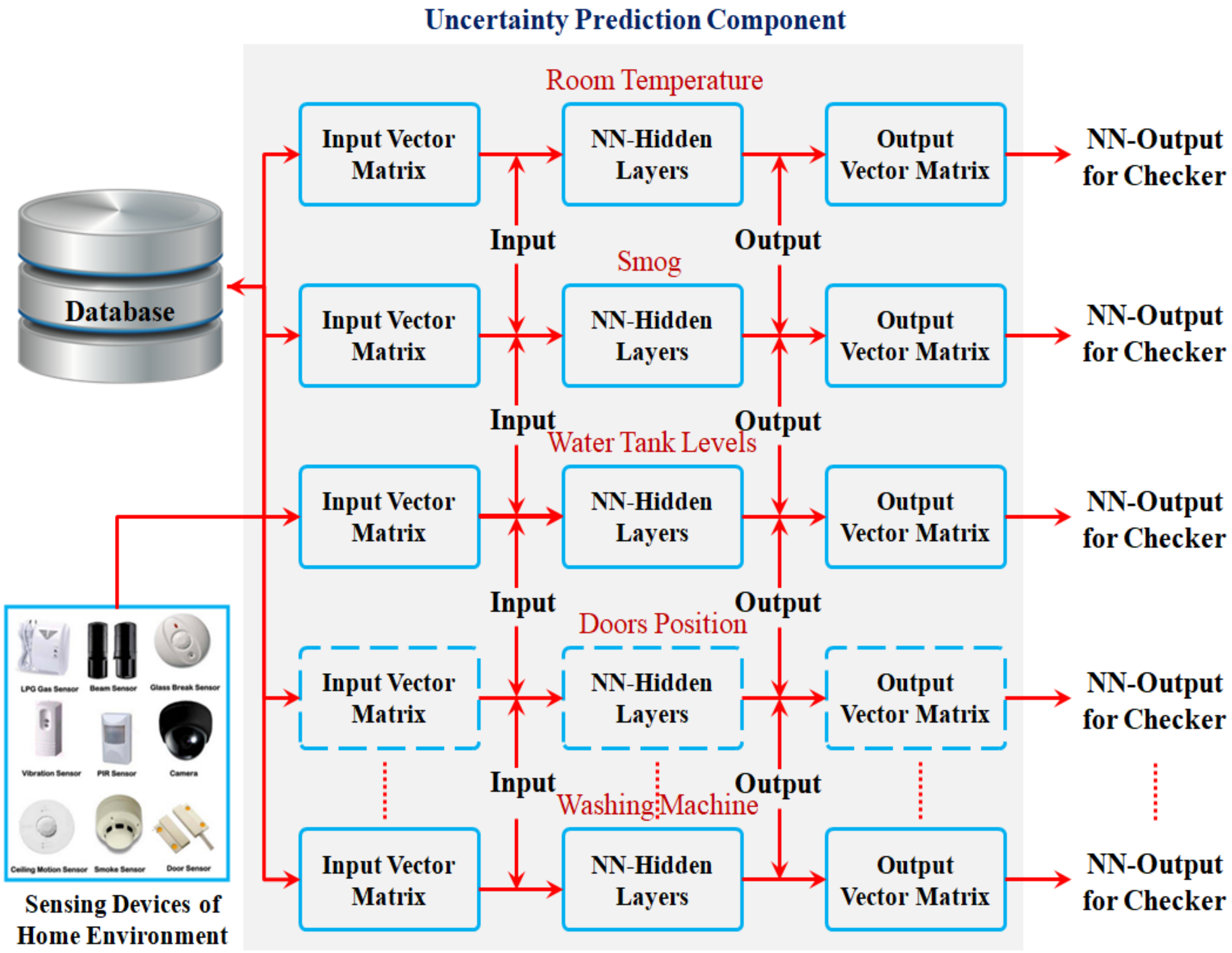


Fig. 1. Different Layers of Neural Network Parameters.

E. The Interaction between the Input Layers

The interaction between different layers is carried out through the hidden layers by using the weights of neurons in terms of ‘0’ and ‘1’. The UPC provides a continuous assessment between different inputs and hidden layer values using the Sigmoid function and is given by

$$Sigmoid\ (x) = \frac{1}{(1+e^{-x})} \tag{2}$$

Firing rules are very flexible to calculate the timing of firing neurons based on the input patterns. However, it is very important to note that the firing rules applied at input and output stages are different; and the generalization of neurons will be considered at both the ends as they are sensible for random patterns applied at the time of training the hidden layers. Such a flexible property of the firing rules is implemented by using the Hamming distance technique [13]. This flexibility helps UPC to work efficiently for finding multiple and complex natured uncertainty accurately.

This system will help the house owners and the builders to create a peaceful environment in the society by avoiding possible uncertain events in their colonies by predicting the unexpected accidents.

F. Datasets of different Activities in Smart Home Environment

In this work, a dataset is created with daily living activities in smart homes, with well-adapted set of actions for a particular constraint. Five goals have been established which need to be met while recording the dataset.

Goal-1: Record the live home environment data and save it to a hard disk.

Goal-2: Establish and classify the realistic routines of the owners from general public (i.e. visitors).

Goal-3: Use a long-time scale for recording the data.

Goal-4: Ensure to connect more number of smart home sensors to more number of appliances and objects for obtaining more accuracy with the end results.

Goal-5: Now label all four primary contexts with accurate dimension throughout the experiment.

At the time of data collection phase three issues may influence seriously as follows: i) some of the appliances are not used by the owners and may use only few parts of the apartment which are more familiar; ii) some of the common activities performed at home may not be followed by everyone due to different cultural and traditional reasons (for example: five times prayer followed by majority of population in Saudi Arabia, but people visiting from different countries may not follow the same pattern); and iii) some of the activity classes may have few instances during data collection phase, and may influence the training neural network with appropriate information may become a challenging task.

III. Proposed Architecture for Smart Home Environment

In the proposed architecture, there are three main components to address: i) Sensing Devices, ii) Runtime Monitoring Component, and iii) Communication System as shown in Fig. 2. All these components are connected by using various embedded networking devices and IoT devices for establishing a communication interface between internal devices and end-users by using a middleware processing included with neural network systems.

1) Sensing devices: The sensing devices are connected with different types of sensors and their interfacing devices using embedded systems and electronic circuitry. Most of them are operated using microprocessors for different types of real-time applications based on sensors and their controlling operations.

The sensors involved in the smart homes/devices are considered to be the eyes and ears of that environment, which help to inform the owners about any kind of uncertain situation. These devices help to avoid major accidents and reduce the level of damages at home environments [15]. Automated devices control the appliances, events, and activities based on the instructions given by the microprocessors/microcontrollers. They help to monitor and react to different changes in the environment in the absence of human beings.

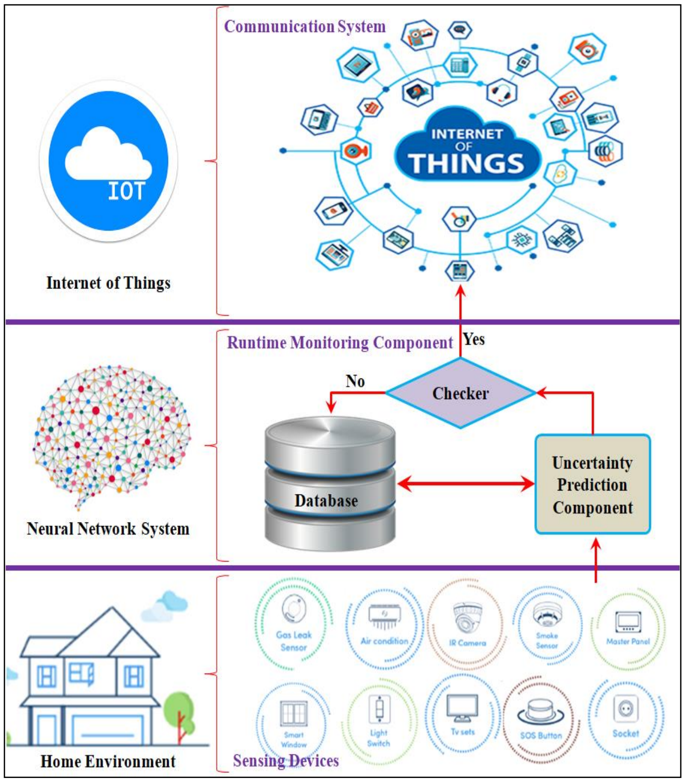


Fig. 2. Proposed NN based Architecture for Smart Home Environment.

2) Runtime monitoring component: The runtime monitoring component in this work includes a database, and checker component along with the uncertainty prediction component, which is completely functioning based on the neural networks-based system. Apart from these activities, the runtime monitoring component also involved with the mitigation process, which helps to assess different scenarios and identify the seriousness of an event taking place in the home environment.

a) Database: It helps to interact with different types of input and output devices to accumulate and store the information. Based on the demand/instructions of the microprocessor, the database helps different components to provide the desired information or store the information by establishing two-way communications. In this work, the database stores the ideal values of different parameters (such as room temperatures, humidity, door positions, motion, water tank levels, gas leakages, etc.) and provide the same to the uncertainty prediction component in the form of a input vector matrix to compare with the latest values at the input nodes of the neural network from the sensing devices. The database design in this work used MySQL application due to simplicity and faster operation as compared to ORACLE or SQL based database designs.

b) Checker: This will check the executed output results of the UPC and ideal values that are stored in the database to identify the environment of the house regularly. If the obtained results are within the controlled range of the ideal values means there is no problem in the home environment. The checker will send an alert signal to the next level (i.e. communication system using IoT devices) if any kind of big difference is observed with the values of UPC and the ideal values stored in the database.

3) Communication system: The communication in this work is established using IoT based system in three ways, i.e. between the targeted objects (various sensors), web servers and the Internet, and different types of social objects (such as mobiles, laptops, internet-based accessories, etc.) [16].

a) Tagged Objects: This category includes RFID, low power sensors, monitoring devices, etc. from the sensing devices to observe the status of the home environment for temperature, water levels, etc. as mentioned above in Fig. 5.

b) Web Servers and Internet: There are so many other technologies for short-range communication using ZigBee, Wi-Fi, etc. [17]. However, they are not capable to communicate by using the low power sensing devices in the IP networks. Therefore, web-based services are needed to get integrated with these sensor nodes for communicating longer distances.

c) Social Objects: Most of the time, data received from such types of monitoring devices will be huge and the relevant executed information will increase exponentially. Therefore, the role of social objects, clouds, etc. plays a vital role in storing the data and using the same effect at the appropriate locations/scenarios. On the other hand, various security issues also can be addressed easily by using such type of social objects, as the services providers will take care of encryption, authentication, and are more cost-effective for the implementation.

To implement this architecture for predicting the uncertainties using NN based runtime monitoring component needs to follow certain rules as discussed in Section-4 and Section-5.

IV. Uncertainty Prediction using the Proposed Method

The list of events taking place using the proposed method are given below as shown in Fig. 3 in a flowchart and some of the highlights are listed below:

- The ideal condition of every device/accessory must be defined by the user in the home environment.
- Based on the inputs the database will help the uncertainty prediction module to generate the appropriate training values for the hidden layers.
- The sensor(s) will try to identify the changes at the home environment regularly and will transfer the same information to the next level, i.e. for UPC and to the runtime monitoring component.
- In the UPC, the weights of the hidden layers will be comparing the input vector matrix and provide the resultant output for the checker component.
- The checker component will compare the ideal values of the database for different specific parameters with the predicted values.
- For any kind of huge gap between the ideal values and the predicted values, an alert will be passed through the IoT section, communicating the owner about the uncertainty.

A. Uncertain Parameters of a Home Environment

In this work, the definition of certain and uncertain conditions of the home environment are listed as shown in Table I. As per the guidelines of the US Environmental protection agency, the pollution levels of smog containing CO of 15.5 ppm to 30.4 ppm and NO_2 of 0.65 ppm to 1.24 ppm [18].

In a certain condition, the people living in the home will feel comfortable and it will be safe from any kind of harmful situations. Whereas, in case of uncertain conditions are very much uncomfortable due to the changes in conditions or any kind of unfortunate events such as weather changes, short circuits, etc. For example, the sudden changes in weather increase the humidity in the room and kids might find it uncomfortable at home in the absence of parents.

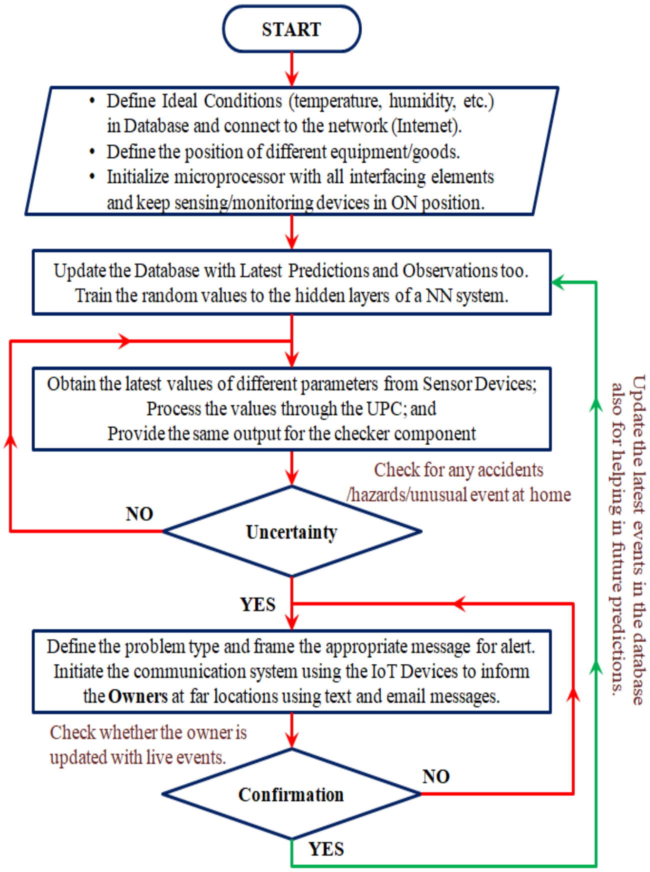

START
• Define Ideal Conditions (temperature, humidity, etc.) in Database and connect to the network (Internet).
• Define the position of different equipment/goods.
• Initialize microprocessor with all interfacing elements and keep sensing/monitoring devices in ON position.
Update the Database with Latest Predictions and Observations too. Train the random values to the hidden layers of a NN system.
Obtain the latest values of different parameters from Sensor Devices; Process the values through the UPC; and Provide the same output for the checker component
Check for any accidents /hazards/unusual event at home
NO
Uncertainty
YES
Define the problem type and frame the appropriate message for alert. Initiate the communication system using the IoT Devices to inform the Owners at far locations using text and email messages.
Check whether the owner is updated with live events.
Confirmation
NO
YES
Update the latest events in the database also for helping in future predictions.

Fig. 3. Flowchart Shows the Events of the Smart Home Environment for Uncertain Events.

TABLE. I. DIFFERENT PARAMETERS WITH THEIR NORMAL AND UNCERTAIN CONDITIONS

Parameters	**Certain**	**Uncertain**
Temperature	17 to 35 ^{0}C	Below/Above 35 ^{0}C
Smog	Less than the uncertainty range.	15.5-30.4 ppm for CO & 0.65-1.24 ppm for NO$_2$
Humidity	30 to 50 percent	Beyond or below the range
Water Levels	Must be around 85%	More than 85%

Similarly, in a smart home, considering human activity recognition (HAR) [19] helps to identify the regular activities performed by the owners and to predict the uncertainties. Some of the common practices followed by the owners at smart home are listed as possible activity classes and are grouped by places in the dataset as follows:

Places	**Activity Classes**
Entrance	Enter, Exit, Cleaning
Kitchen	Prepare, Cook, Dish Washing, Cleaning
Living Room	Eat, Watch TV, Using Computer, Cleaning
Toilet	Using, Cleaning
Staircase	Going Up, Going Down, Cleaning
Bathroom	Use Sink, Toilet, Shower, Cleaning
Office	Using Computer, Watch TV, Cleaning
Bedroom	Dressing, Read, Nap, Cleaning

It is very complicated task to consider an occupant following accurate timings on daily basis and when it comes to multi-owners it will become a serious and complex issue to analyse the number of instances using HAR. Berlemont *et al.* suggested a method to deal with the problems related with gestures and actions using Siamese NN (SNN), which represents two identical ANNs runs simultaneously on two different types of instances [20]. It is also important to note that the data types in the dataset will change based on the type of sensing device used for a particular activity. Some of the data may be in binary form (such as door closed or opened, i.e. assuming Close with '0' and Open with '1'). Similarly, data can be an integer, real number, and could be categorical too. The neural network may need to learn the number of instances and their frequency during working days and weekends/holidays. For such complicated information the network must be well trained with maximum activities and their instances. For example, using toilet may have few instances but using staircase and computing at home may have more instances during a weekend/holiday as compared to any working day.

V. DISCUSSIONS

Place-based activity recognition is popularly known approach with primary context dimensions included with identity, time, place and activity. These four dimensions are strongly inter-related with each other. For example, an occupant sleeping in the bed room at the night time considers all four activities: occupant (identity), sleeping (activity), bed room (place), and night time (time) covering all four activities. Consider all sensors in the smart home are represented by $S = \{S_1,\ S_2, \dots\ S_n\}$ and all the activity classes are represented by $\boldsymbol{\mathcal{A}}$ and current instances with a, which is a subset of $a \in \boldsymbol{\mathcal{A}}$. Combining all these may be represented by *global approaches*. Now considering i representing a particular place (i.e. a *local approach*) and set of activity classes are represented by $\boldsymbol{\mathcal{A}}^{(i)}$. Let's say $S^{(i)} = \left\{S_1^{(i)},\ S_2^{(i)}, \dots\ S_{n^i}^{(i)}\right\}$ representing the complete set of sensors at i^{th} place of the smart home. Any local approach at the i^{th} place belongs to a classifier needs to get recognized and the current instance at any point is $a \in \boldsymbol{\mathcal{A}}^{(i)} \cup \{none\}$ with the combined data produced by $S^{(i)}$. Here, $none$ is a dummy class and it is very much needed to assign a meaningful label for an instance unless it is a part of $S^{(i)}$. An activity recognition approach is proposed in this work by learning local models for each instance, place and active classes using the following steps: First, use the localization algorithm to locate an occupant at i^{th} place; and then to recognize the occupant activity at i^{th} place using the local model. Now by considering set of all places of a home by $\mathcal{P}$ and the cardinal of this set is represented by $|\mathcal{P}|$, respective set of decisions are computed for $\left\{\Delta^{(1)}, \Delta^{(2)} \dots \Delta^{(\mathcal{P})}\right\}$.

here,

$$\Delta^{(i)} = \left\{\delta_{1,1}^{(i)}, \dots, \delta_{1,|A^{(1)}|}^{(i)}, \dots, \delta_{|\mathcal{P}|,1}^{(i)}, \dots, \delta_{|\mathcal{P}|,\boldsymbol{\mathcal{A}}^{(|\mathcal{P}|)}}^{(i)}, \delta_{none}^{(i)}\right\} \qquad (3)$$

and $\delta_{k,j}^{(i)} \in [0,1]$ represents degree of membership with the decision of a classifier at i^{th} place of j^{th} activity class and k^{th}

place. In the proposed approach the local models will learn to recognize different activity classes which occurs at respective places, i.e. if $i \neq k$ gives $\delta_{k,j}^{(i)} = 0$. Here the NN needs to make a decision and the decision fusion step is to compute set of fused decision ($\bar{\Delta}$) is given by

$$\bar{\Delta}= \left\{\bar{\delta}_1^{(i)}, \dots, \bar{\delta}_{|A^{(1)}|}^{(1)}, \dots, \bar{\delta}_1^{(|\mathcal{P}|)}, \dots, \bar{\delta}_{\mathcal{A}^{(|\mathcal{P}|)}}^{(|\mathcal{P}|)}, \bar{\delta}_{none}\right\} \tag{4}$$

Now from each set of decisions $\{\Delta^{(1)}, \dots, \Delta^{(\mathcal{P})}\}$ each place is considered from a smart home and from $\bar{\Delta}$ a conclusion is made to identify the class of a particular instance at a^{th} activity from p^{th} place when a class is having maximum decisions in $\bar{\Delta}$. Therefore,

$$\bar{\delta}_a^{(p)} = \max\left(\begin{matrix} max \\ j, k \end{matrix}\, \bar{\delta}_j^{(k)}, \bar{\delta}_{none}\right) \tag{5}$$

From the above Eq. (5), a given instance will be related to only one activity when the similar activities are occurring throughout the home simultaneously. By using maximum values of $\bar{\Delta}$, modification of decision process for multiple activity labels is possible. The proposed place-based activity recognition scheme is shown in Fig. 4.

There is a requirement of expert knowledge for setting up entire process in a smart home, and need to define the membership of each sensor with respective place is essential. Sensors are generally installed closely with each other and the possibility of discovering same places by the sensors is more. To tackle/address this problem, clustering methods are implemented. The same data collected by nearest sensors helps to find correlated data. It is necessary to standardize the sensor data in a way to have 0 mean and a variance of 1. The training dataset is given by considering mean as $\bar{s}$ and standard deviation of outputted values the sensor with σ as

$$s_t' = \frac{s_t - \bar{s}}{\sigma} \tag{6}$$

Also considering the noise parameters in the collected data, it is essential to ensure a noise reduction process is being carried out. A basic filtering method is used to control the noise and it is controlled by $\beta \, \epsilon \, [0,1]$ is given by

$$s_t' = \beta s_t + (1 - \beta) s_{t-1} \tag{7}$$

Here maximum filtering process is considered when $\beta = 0$; minimum filtering also considered when $\beta = 1$ when $s_t' = s_t$ for any value of t. The basic method applied over raw sensor data is shown in Fig. 5 with the resultant filtered output, where the original signal shape is conserved by softening all the noisy oscillations.

A. Classical Classifiers

There are several classifiers which can be used for predicting the uncertainties in a smart home environment. These classifiers included with multi-layer perceptrons (MLP) of ANN, Bayesian networks (BN) used for probabilistic graphical modelling, support vector machines (SVM) are used for different types of kernel methods, dynamic time warping (DTW) is used for measuring the geometric similarities, and hidden Markov models (HMM).

MLP uses the feed forward techniques of ANN with no loops and in this method the output is connected with the input of another neuron at the next stage or layer. By considering $n_i^{(j)}$ as i^{th} neuron of j layer with the inputs from a set of outputs neurons of N_{j-1} from previous layers as $\left\{y_i^{(j-1)}, \dots, y_{N_{j-1}}^{(j-1)}\right\}$gives an output say $y_i^{(j)}$. This output can be computed by using the following equation:

$$y_i^{(j)} = \varphi\left(b_i^{(j)} + \sum_{k=1}^{N_{j-1}} w_{k,i}^{(j-1)}, y_k^{(j-1)}\right) \tag{8}$$

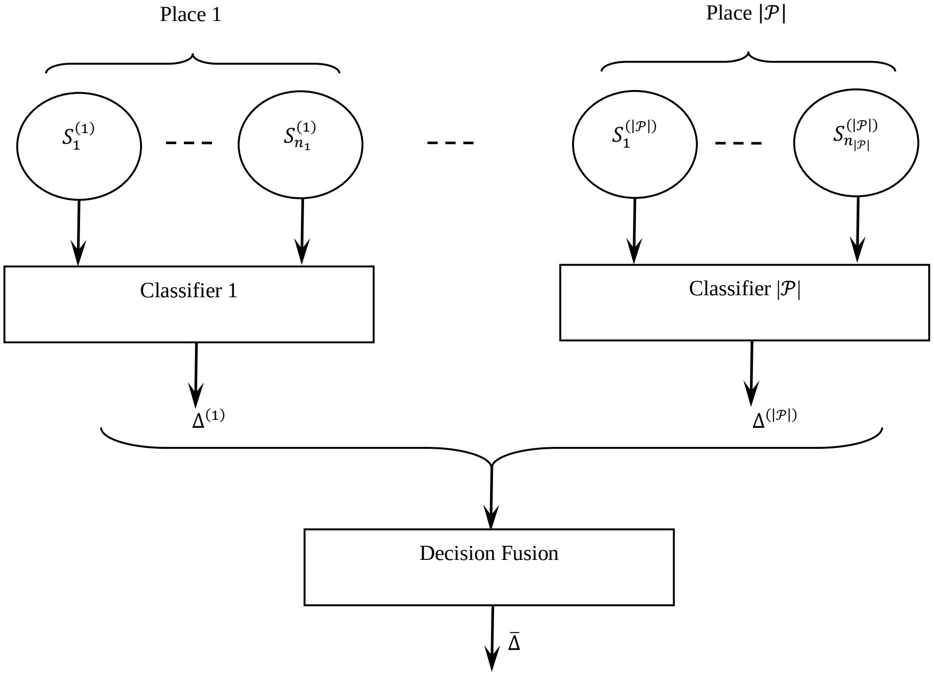


Fig. 4. Shows the Place-based Activity Recognition Scheme.

here, $b_i^{(j)}$ is the bias neuron, $w_{k,i}^{(j-1)}$ is the weight between two neurons (i.e. between k and $j-1$; and i and j), and finally activation function of neuron is represented by φ, which is equal to sigmoid function as listed in Eq. 2.

SVMs are the generalization of linear classifiers with two classes. For a vector of the input data $x = (x_1, \ldots, x_N)$ a constructing linear classifier is not possible with the help of vector weights $w = (w_1, \ldots, w_N)$. Therefore, the output of SVMs can be given by

$$y = wx^T + w_N \tag{9}$$

If the value of $y \geq 1$ it is classified in class 1, and if $y \leq -1$ then it is classified in class 2.

BN is considered as a directed acyclic graph representing joint probability distribution of different variables forming as a set. For a five variable inputs of $\boldsymbol{x} = (x_1, \ldots, x_5)$ the BN is represented by different conditional dependencies as shown below:

$$p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4)p(x_5|x_2, x_3, x_4) \tag{10}$$

For the purpose of activity classification, any one variable of BN represents the activity class and remaining variables represents the obtained sensor data. For example, in the following Eq. 10 activity class is represented by x_1 and sensor data is represented by x_2, x_3, x_4, x_5.

$$\underset{x_1}{argmax}\ p(x_1|x_2, x_3, x_4, x_5) \tag{11}$$

There are many approaches existing to learn from the trained data for the BN structures and conditional probabilities of BN. There are some complex learning problems such as structure learning as compared to distribution learning.

In case of DTW, values of two time-dependent sequences are compared in a geometric approach. For example, assume values of two sensors as x_t and $y_{t'}$,with a time steps of $t \in [\![1, T]\!]$ and $t' \in [\![1, T']\!]$. To compare the sensor values it is necessary to conduct cost measurement with $c: [\![1, T]\!] \times [\![1, T']\!] \rightarrow [0, +\infty]$. Generally, for *similar* values of x_t and $y_{t'}$, output of cost function will be smaller values; and for *dissimilar* values of x_t and $y_{t'}$,output of cost function will have larger values. The warping path is defined as a sequence $p = (p_1, \ldots, p_n)$ where the value of $p_i = (t_i, t_{i'}) \in [\![1, T]\!] \times [\![1, T']\!]$. This warping path need to consider and respect two conditions: boundary conditions ($p_1 = (1,1)\ and\ p_n = (T, T')$); and step size conditions ($\forall_i [\![1, n-1]\!], p_{i+1} - p_i \in \{(1,0), (0,1), (1,1)\}$). The advantage of boundary conditions is to ensure both start and end points are aligned together. Whereas step size conditions enforces all elements from both the sequences are part of warping path for at least once with none of the duplicate pairs in p. The later step also enforces monotonicity of the path (i.e. $t_i + t_{i+1}$and $t'_i \leq t'_{i+1}$. Finally, the cost c_p of warping path p is given by

$$c_p(x, y) = \sum_{i=1}^{n} c\left(x_{t_i}, y_{t'_i}\right) \tag{12}$$

Therefore, DTW distance between two sequences is representing the overall cost of optimal warping path, which is smallest of all possible costs among different warping paths.

B. *Experiment Results*

At the time of processing data missing values are replaced with cubic spline interpolation and a low-pass filter (with $\beta = 0.1$) was used for noise reduction. Similarly, care was taken to ensure the value of mean is ‘0’ and standard deviation of the sensors is equal to 1. Sensor values as input vectors are converted to feature vectors by resampling each instance with 20 time steps. WEKA library was used to implement the MLPs, SVMs, and BNs in this experiment [21].

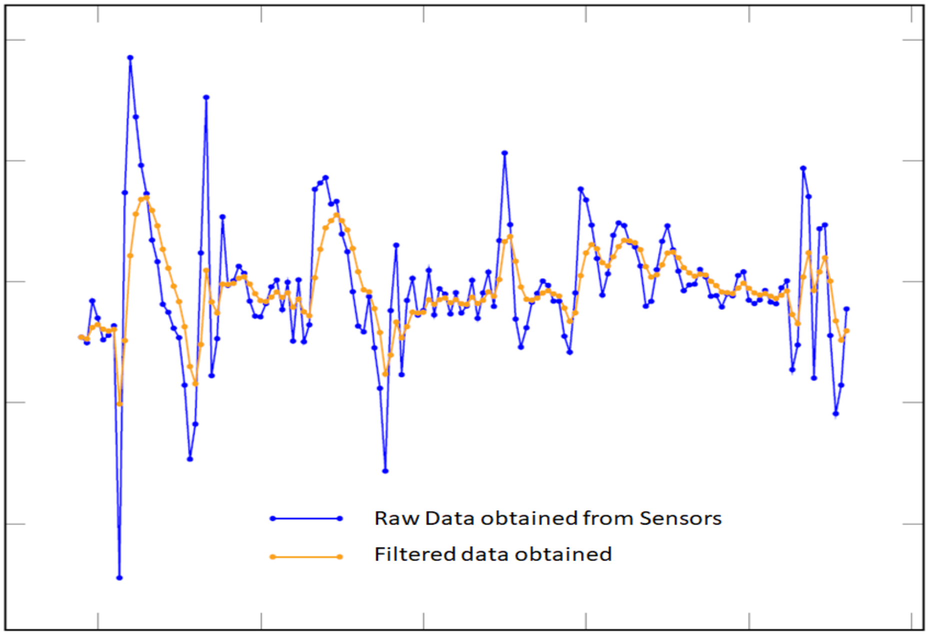


Fig. 5. Shows the Filtered Data with β = 0.3 after Compiling the Same with Raw Data Obtained from the Sensors.

Activity recognition performance of this work is compared with place-based approach with the global approach by using different types of classifier types and decision fusion methods. It is also important to note that most of the classes are not frequent in any smart home environment and the frequency may vary from class to class. Weighted F$_1$ score is used to compare the performances at different levels, rather comparing only the accuracy part of it. Performance measure helps to provide equal weights for all classes, irrespective of the total number of instances they consists off. For example, consider a set with activity instances (χ_a) with true label a for all activity classes $a \in \boldsymbol{\mathcal{A}}$, the F$_1$ score is calculated as shown below:

$$F_1 = \sum_{a\in\boldsymbol{\mathcal{A}}} \frac{2}{|\chi_a|} \cdot \frac{Precision(a).Recall\ (a)}{Precision(a)+Recall\ (a)} \qquad (13)$$

here, the Precision (a) is a ratio of number of correctly classified instances as a to the total number of classified instances as a. Similarly, Recall (a) is the ratio of number of correctly classified instances as a to the total number of instances of class a.

In the opportunity dataset there are about 80 classes of actions are labelled in this work and an additional class with label 'None' is placed considering a case when no action is performed. In this place-based approach the sensors are completely based on the locations of action classes as mentioned in the dataset in Section 4. In this work, five places (as shown in Fig. 6) are identified for experimental environment where recording of Opportunity dataset is possible: in the Kitchen, Tables, Bed Rooms, Exits, and TV's. All the sensors and action classes are distributed among these 5 places based on the number of classes. For example, in a kitchen there are 10 drawers (Open and Close), Dishwasher (ON and OFF), Lights (ON and OFF), Fridge (Open and Close), Stove (ON and OFF), and Windows (Open and Close) gives a total 30 action classes. Similarly, for three bedrooms there are 18 action classes by considering light, fan and air condition (AC) as ON/OFF. On the table, it is considered to have 12 action classed with the in-built sensors to some of the objects placed on a table. There are four TV's with 8 action classes and six exit (E$_1$ to E$_6$) gives a total of 12 action classes.

The evaluation process in this approach used 10-fold random cross-validation. The five sets of instances were used to train the neural network. The performance of this approach was evaluated by using the test set which never existed before the training the hidden layers. For selecting parameterization of the classifiers/decision fusion validation set was used. Evaluation of validation set took at first place with the learned models and this helps to avoid any kind of bias at the time of evaluation phase. F1 scores of five different places using the four classifiers are given in Table II from the Opportunity dataset.

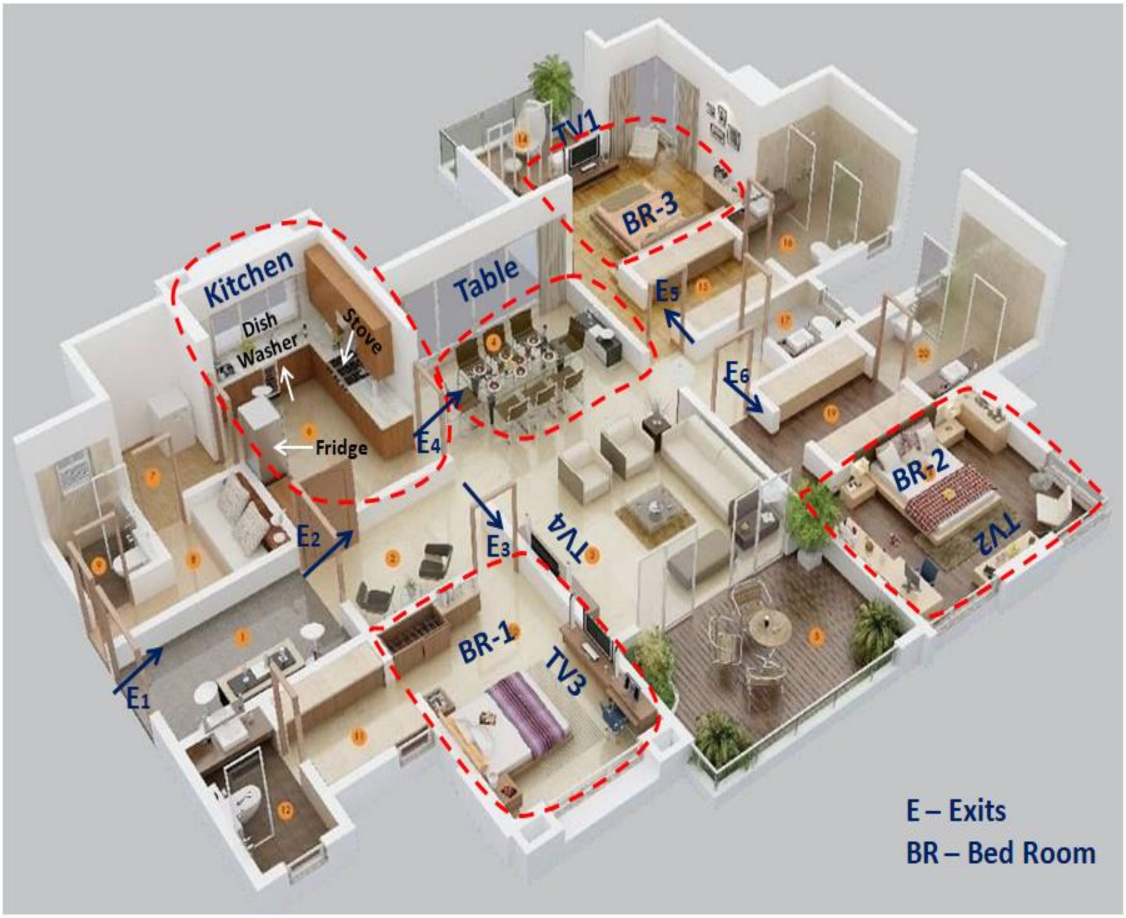


Fig. 6. Shows the Five Places where the Opportunity Dataset was Recorded from different Sensors from the Smart House Environment: Table, Bedrooms, Kitchen, Exits and TV's.

From the Table II, it can be seen that the performances are depending on different types of classifiers. Performance using DTW shows poor results as compared BN at the place Kitchen. The observations also show a huge gap between some of the classification performances at some places. Overall recognizing actions from Kitchen seem more complicated as compared to other classifiers. Now in a process of comparing place-based and global approaches, all the places modelled by using same classifier as compared to the global approach in which all classifier types were used for considering a global model action from Opportunity dataset. From the results it can be seen that the place-based approach is far better than the global approach as shown in Table III.

The gap between these two approaches is not significant statistically because the F1 score deviation of the classifiers has a big gap when it is observed in MLP and DTW from the Table III. Also note that the standard deviation of all classifiers is small for place-based approach as compared with global approach, which is due to decision fusion step that tends to go for average out of the overall results.

In this work, anticipation is also made to calculate the computing times (see Table IV) as a benefit of place-based approach. The time taken to run a training phase and activity recognition is observed to be faster in place-based approach when compared with global approach. To calculate the computational times, it is necessary to use a high-end system configuration consisting of high frequency of operation and huge RAM. Here, the computing times are considered for only three classifiers (MLP, SVM, and BN; and ignored DTW since it is very slow as compared to other three classifier types as seen from in Table II) along with global approach over the Opportunity dataset. Whereas in case of decision fusion step it was being ignored due to very slow computing times and can be neglected as compared to any computing times.

TABLE. II. F1 SCORES FOR DIFFERENT CLASSIFIERS FROM FIVE PLACES IN OPPORTUNITY DATASET

	Place				
Classifier	**Kitchen**	**Table**	**Exits**	**Bedrooms**	**TV's**
MLP	**94.08%±1.67%**[1]	**98.57%±0.43%**[1]	99.12%±0.43%[1]	**96.62%±1.43%**[1]	**99.02%±1.03%**[1]
SVM	93.87%±1.23%[2]	98.47%±1.63%[2]	**99.34%±0.36%**[2]	95.71%±1.53%[2]	98.87%±1.73%[2]
BN	91.54%±1.29%	98.37%±0.43%	98.91%±0.53%	94.92%±1.23%	98.59%±1.68%
DTW	84.67%±1.13%	98.23%±0.37%	98.76%±0.93%	94.33%±1.58%	98.42%±1.36%
Parameters	[1] 100 hidden neurons, 120 epochs, 0.2 learning rate, and 0.1 momentum [2] C = 1000, γ = 0.01				

TABLE. III. F1 SCORES FOR DIFFERENT CLASSIFIERS USING GLOBAL APPROACH AND/OR PLACE-BASED APPROACHES FROM THE OPPORTUNITY DATASET

	Classifier			
Approach	**MLP**	**SVM**	**BN**	**DTW**
Global	90.11%±1.57%[1]	90.12%±1.03%[2]	**90.72%±1.43%**[1]	79.62%±2.43%[3]
Place-based	**93.42%±1.36%**[3]	**91.23%±1.25%**[4]	89.36%±1.83%[5]	**84.33%±1.38%**[6]

Parameters used for global approach
[1] 100 hidden neurons, 120 epochs, 0.2 learning rate, and 0.1 momentum
[2] C = 1000, γ = 0.003
Decision fusion used for place-based approach
[3] SVM stacking using C = 100, γ = 0.01
[4] MLPstacking using 120 hidden neurons, 120 epochs, 0.2 learning rate, and 0.1 momentum
[5] MLPstacking using 60 hidden neurons, 120 epochs, 0.2 learning rate, and 0.1 momentum
[6] SVM stacking using C = 20, γ = 0.05

TABLE. IV. AVERAGE COMPUTING TIMES FOR DIFFERENT CLASSIFIERS FROM FIVE PLACES IN OPPORTUNITY DATASET ALONG WITH GLOBAL APPROACH

		Place					
Classifier	**Phase**	**Kitchen**	**Table**	**Exits**	**Bedrooms**	**TV's**	**Global**
MLP	Training Test	956.36±159.32 13.68±1.32	714.11±110.05 14.68±1.67	689.85±19.92 11.24±1.54	547.96±95.69 10.98±1.27	521.63±58.25 9.58±1.25	15,586.36±1599.3 26.98±1.68
SVM	Training Test	29.65±0.96 7.58±0.08	24.15±0.32 9.69±0.18	19.64±0.32 7. 81±0.29	18.06±0.72 9.58±0.42	16.38±0.32 8.09±0.95	39.68±0.75 28.18±0.58
BN	Training Test	21.37±0.39 9.68±0.17	15.61±0.85 8.81±0.45	14.08±0.38 7.31±0.21	11.58±0.24 6.85±0.37	9.24±0.75 6.28±0.11	29.67±0.62 13.18±0.01
DTW	Training Test	0 4215.81±283.72	0 3512.90±224.29	0 3213.61±191.76	0 2973.58±147.57	0 2513.84±1381.48	0 7214.06±389.24
Computing Time in *Seconds*							

Considering the proposed place-based approach executed with multi-core computer processor, training phase can be parallelized. That means all places can be computed simultaneously by assuming decision fusion takes negligible time. Due to this approach the overall computing times of proposed place-based approach is shorter as compared with the global approach for both training and testing phases. For example, considering the MLP for any place as compared with the global approach the training time and testing times are shorter in place-based approach (as 15,586.36 seconds is bigger than any values of MLP classifier).

VI. Conclusions

In this work, a unique method to predict the uncertain events at the home environment is proposed using a neural network-based system. As compared to the earlier home automation systems, the present method can detect the possible uncertain situations and hazards in an efficient way to improve the safety measures for the people living in the home from any kind of serious life and property damage. There is a lot of improvement in monitoring the events and activities using the sensors and monitoring devices effectively adding the risk prediction component. Both of them working simultaneously and the UPC component will try to judge and mitigate the events based on the previous data also in the absence of the immediate present data when there is a problem with sensors, such as power failures, technical glitches, etc. The place-based approach results proved to be very much useful for the prediction of uncertainties in the smart home environment. Four classifiers were used to examine the performance of proposed UPC and MLP proved to be more accurate in terms or predicting uncertainties. Place-based approach delivered better results as compared with the global approach and the training and testing times required by both the approaches also shown similar results.

Acknowledgment

The authors gratefully acknowledge the support and facilities provided by the Management and Department of Computer Science and Information, University of Hail, Hail, Saudi Arabia.

References

[1] E. Park, A. del Pobil, and S. Kwon. "The role of internet of things (IoT) in smart cities: Technology roadmap-oriented approaches." Sustainability 10, no. 5 (2018): 1388.

[2] A. Reyna, C. Martín, J. Chen, E. Soler, and M. Díaz. "On blockchain and its integration with IoT. Challenges and opportunities." Future Generation Computer Systems 88 (2018): 173-190.

[3] X. Yuan, and S. Peng. "A research on secure smart home based on the internet of things." In 2012 IEEE International Conference on Information Science and Technology, pp. 737-740. IEEE, 2012.

[4] V. Miori, and D. Russo. "Anticipating health hazards through an ontology-based, IoT domotic environment." In 2012 Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, pp. 745-750. IEEE, 2012.

[5] M. Jung, J. Weidinger, W. Kastner, and A. Olivieri. "Building automation and smart cities: An integration approach based on a service-oriented architecture." In 2013 27th International Conference on Advanced Information Networking and Applications Workshops, pp. 1361-1367. IEEE, 2013.

[6] R. Piyare. "Internet of things: ubiquitous home control and monitoring system using android based smart phone." International journal of Internet of Things 2, no. 1 (2013): 5-11.

[7] A. Jacobsson, M. Boldt, and B. Carlsson. "On the risk exposure of smart home automation systems." In 2014 International Conference on Future Internet of Things and Cloud, pp. 183-190. IEEE, 2014.

[8] G. V. Vivek, and M. P. Sunil. "Enabling IOT services using WIFI-ZigBee gateway for a home automation system." In 2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), pp. 77-80. IEEE, 2015.

[9] V. Puri, and A. Nayyar. "Real time smart home automation based on PIC microcontroller, Bluetooth and Android technology." In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1478-1484. IEEE, 2016.

[10] S. Pirbhulal, H. Zhang, M. E Alahi, H. Ghayvat, S. Mukhopadhyay, Y. T. Zhang, and Wanqing Wu. "A novel secure IoT-based smart home automation system using a wireless sensor network." Sensors17, no. 1 (2017): 69.

[11] J. Saha, A. K. Saha, A. Chatterjee, S. Agrawal, A. Saha, A. Kar, and H. N. Saha. "Advanced IOT based combined remote health monitoring, home automation and alarm system." In 2018 IEEE 8th annual computing and communication workshop and conference (CCWC), pp. 602-606. IEEE, 2018.

[12] O. Hamdan, H. Shanableh, I. Zaki, A. R. Al-Ali, and T. Shanableh. "IoT-based interactive dual mode smart home automation." In 2019 IEEE International Conference on Consumer Electronics (ICCE), pp. 1-2. IEEE, 2019.

[13] A. Jadi. "An Early Warning System for Risk Management." Ph.D. Thesis. Software Technology Research Laboratory, DeMontfort University, 2013.

[14] A. Jadi, H. Zedan, and T. Alghamdi. "Risk management based early warning system for healthcare industry." In 2013 International Conference on Computer Medical Applications (ICCMA), pp. 1-6. IEEE, 2013.

[15] M. Yerragolla, K. Pallela, and I. P. Gera. "Intelligent security system for residential and industrial automation." In 2016 IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics Engineering (UPCON), pp. 229-234. IEEE, 2016.

[16] A. Singh, D. Kumar, and J. Hötzel. "IoT Based information and communication system for enhancing underground mines safety and productivity: Genesis, taxonomy and open issues." Ad Hoc Networks 78 (2018): 115-129.

[17] S. Al-Sarawi, M. Anbar, K. Alieyan, and M. Alzubaidi. "Internet of Things (IoT) communication protocols." In 2017 8th International conference on information technology (ICIT), pp. 685-690. IEEE, 2017.

[18] T. Fitz-Simons. Guideline for reporting of daily air quality: Air Quality Index (AQI). No. PB-99-169237/XAB; EPA-454/R-99/010. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, NC (United States), 1999.

[19] J. Cumin. "Recognizing and predicting activities in smart homes." PhD diss., 2018.

[20] S. Berlemont, G. Lefebvre, S. Duffner, and C. Garcia. "Class-balanced siamese neural networks." Neurocomputing 273 (2018): 47-56.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter 11, no. 1 (2009): 10-18.

“Onto-Computer-Project”, a Computer Project Domain Ontology : Construction and Validation

Mejri Lassaad[1]
College of Computer and Information Science
Jouf University, Saudi Arabia

Hanafi Raja[2], Henda Hajjami Ben Ghezala[3]
National School of Computer
University of Manouba Manouba, Tunisia

Abstract*—Ontologies, nowadays, play a primordial role in the representation, the re-use and the sharing of knowledge of a well given domain in a consensual and explicit way more precisely in the computing field. It is in this context that we have proposed a domain ontology baptised onto-computer-project which presents the key to our research goal. This essential goal is to arrive at a final step to elaborate a knowledge based system for computer projects reusing. The aimed system is essentially based on the construction of a memory projects. This memory projects could be defined as a collection of historical and achieved projects around the sphere of computer. This sphere is so wide including many subfields beginning at the database, software engineering fields and arriving at the fields of artificial intelligence, computer vision and so on. This research work requires at first to construct a well-defined ontology in the way to structure and to unify vocabulary often shared by multi actors in the domain of computer projects. To concretize this goal, our paper will describe a construction approach for the proposed domain ontology which is mainly based on an existing methodology named “methontology”. The proposed ontology construction approach, which is composed of seven steps, is the result of a comparative study between some ontology construction approaches belonging to different categories of methodology. In fact, we can distinguish four main categories of ontology development approaches: ontology construction approaches from zero, text-based construction approaches, building approaches based on the reuse of already existing ontologies, and crowd souring-based approaches. In our research work, we are interested by the approach of building ontology from zero. Indeed, the construction of the proposed ontology follows an autonomous approach which is not based on any other existing ontology or the updating of an already constructed ontology. In addition, in this paper we are interested by the problem of validating the content of domain ontology and in this context we have proposed an incremental approach for validating the proposed ontology which is composed of six steps. In this context we have studied some ontology validation approaches: those which are questionnaire based, others based on question answering. The problem here that all approaches studied are single actor approaches where a single validation actor can validate the entire ontology and this by applying the semantic and the structural validation definitively with no return. The main originality of our validation approach consists essentially of three criteria: the incremental validation, the multi-intervention, and the respecting of the “V” cycle. In fact, the passage from one validation step to another results in an update of the initial ontology and this by the intervention of three experts (project management expert, a project computer expert and a specialist in ontology engineering). Our proposal approach requires a feedback between all the validation phases and can return to any expert for revalidation if needed. The result of this research is improved a validated ontology which is allowed us to build our project memory and to feed our knowledge base which will serve us to develop our knowledge-based system.

Keywords*—Domain ontology; ontology construction; ontology validation; computer project; project memory; knowledge representation

I. Introduction

Ontologies now play a major role in the representation and modeling of knowledge. Their main objective is to formalize the knowledge of a domain and thus add a semantic layer to computer systems and applications. In addition, the development of a new ontology makes it possible to explicitly represent the knowledge of a domain by means of a formal language, in order to be able to be manipulated automatically and shared easily [1].

Indeed, ontology consists of a set of concepts organized using hierarchical and specialized relationships representing a means of expression, sharing and reuse of knowledge, usable by all actors. In addition, an ontology is a computer artifact conceptually modeling knowledge, an indexing system for a specialty area, a theory of scientific content, a representation of shared knowledge or a modeling of reality [2]. It is in this context that we have exploited the proposal of a computer domain ontology to achieve our main research goal which is the capitalization of knowledge from computer project memory. A project memory can be defined as an explicit and continuous representation of knowledge, data or data source within an organization which contains the context in which the knowledge has been created [3]. Therefore, project memory allow professional actors to reuse and share knowledge, which has been capitalized from previous projects in order to carry out a new one [4].

In other hand, the newly created ontology must be validated and evaluated thanks to either experts or standard validation tools. So, we can identify two scenarios [13] which justify the validation of the ontology: an adequate ontology will allow better reuse of the data and ontologists need methods to evaluate and validate their models in order to encourage them to share with confidence their results with the community. It is in this context that this paper will focus primarily on these two aspects: the choice of the methodology of construction of our ontology and the proposition of a validation approach of the proposed ontology.

The paper is organized as follows. After the introduction, Section 2 consists of the state of the art which is composed of two sub-sections: In the first Sub-section the ontology construction methodologies are reviewed then a comparative study between these methodologies is introduced. In Section 2, we will describe both the main related works of ontology validation approaches and a discussion study. Section 3 presents an ontological construction approach and Section 4 describe the incremental validation approach for our proposed ontology. Finally Section 5 reveals the main conclusion and future works.

II. Related Works

This section consists essentially of two parts: The first one reveals a state of art on the methods of ontological construction and a comparative study of the invoked methodologies. The second one describes the main existing works in the literature and a comparative study of ontology validation approaches. In the following, we introduce the major works in the literature associated with the ontology construction methodologies.

A. Ontology Construction Methodology

In ontology engineering, the choice of methods, techniques and tools for the ontology construction process is a key step. Indeed, several methodological approaches have been proposed [6][7] to guide this process. We can distinguish four main categories of ontology development approaches: ontology construction approaches from zero [6], text-based construction approaches [8], building approaches based on the reuse of already existing ontologies [10] and crowd souring-based approaches [9].

In our research work, we will be interested in the approach of building ontology from zero. Indeed, the construction of the proposed ontology follows an autonomous approach which is not based on any other existing ontology or the updating of an already constructed ontology. Moreover, the knowledge and skills defining the essential components of the proposed ontology did not come from textual resources.

For all the reasons mentioned above, we found ourselves obliged to adapt the construction approaches from zero for the development of our domain ontology. In the following, we introduce the major works in the literature associated with this kind of approach.

B. Main Approaches from Zero Description

Several works in the literature are oriented towards this type of approach in what follows we have discussed some proposals.

1) Two-step methodology [11]: As its name indicates, this methodology is composed of two stages: The knowledge organization and the knowledge acquisition and reuse that allow the users collaboratively producing and consuming the knowledge. In the beginning, a Core Reference Ontology [CRO] describing the generic concepts and relations according to the formalized requirements is identified. Then, a Domain Specific Ontology [DSO] is specialized. Only two steps are not enough to describe a complete construction processes. In fact, this methodology is neither documented nor evaluated.

2) On-to-knowledge methodology [OTKM][12]: It is a methodology based on acquired experiences of business activities. It is composed of four stages from identification, to documentation [11]. The stages are given implicitly and the activities are few detailed (general description of the steps and no precision in the choice of components)[3].

3) The methodology proposed by fox & al [13]: This methodology is used in the context of the TOVE project (Toronto Virtual Enterprise). The application of this methodology is motivated by problems which are formulated under form of informal questions that ontology should answer. This methodology has made it possible to develop complex projects in the field of business but remains limited because neither the different stages nor the techniques used are precisely described.

4) OntoDI methodology [14]: This methodology is an ontology development method which has been developed for the implementation of data integration called ontology development on the data integration domain (OntoDI). The main objective of this ontology construction method is the development of knowledge in ontology to manage the problems of semantic aspects in order to support the implementation of data integration. OntoDI has three main phases: pre-development, main development and post-development. And in each part contains several phases.

5) The methontology [15][16]: It is the most widely used methodology in literature. It is the most adopted construction approach for much ontology in different fields. In fact, this method is highly-precised. “Methontolgy” can be applied in all areas, and it can be applied in scratch or text approach.

In order to adopt an approach to construct our ontology, we will propose a comparative study between the methods already mentioned in the previous sub-section.

C. Comparative Study

The comparative study [Table I] is based on four criteria: these criteria are selected in accordance with domain experts.

- Process step: this criterion describes the form or the way in which the construction process steps is defined: detailed, little detailed, very detailed.
- Level of precision: the precision in the choice of the terms, relations and classes during the construction stages. This criterion differs from one method to another.
- Application domain: It serves to know in which domain this method has been applied.
- Type of activity: each process of construction is composed of a set of tasks or activities. Here, we have tried to determine the type of activity.

TABLE. I. COMPARATIVE STUDY BETWEEN CONSTRUCTION METHODS

	Process step	Level of precision	Application domain	Types of activity
Two-step Methodology [11]	Only 2 steps + general description	Generic concept + core ontology	Knowledge management	Formalisation
OTKM methodology [12]	4 steps+ detailed description	Absence of precision	Business Domain	Documentation evaluation
Fox and al 's Methodology [13]	4 steps Not detailed	Lack of precision	Business Domain	Evaluation activity
OntoDI Methodology [14]	6 steps detailed +complex(	Good level of Precision	Data integration	Evaluation +validation
Methontolgy [15]	Detailed description	Good level of Precision	Several domains (FIPA)	Activity (project management support formalization

This comparative study results in the choice of the "Methontology" as a methodology for ontology elaboration. Indeed, "Methontology" is the most precised of all the previous methodologies. In addition, this methodology offers several types of activities and among these activities we mention project management.

D. Ontology Validation Approaches

The ontology validation is a very essential phase in the ontology construction process. Without this step, the ontology could not be exploitable or applicable. This stage becomes more and more complex with the increasing size of ontologies and the use of semantic construction [17].

Several approaches have been proposed in the literature for the validation of over ontology from different applications. In this section we will discuss the main approach of that have been proposed for the ontology validation.

1) Main approaches of ontolygy validation

a) An approach for validating the content of an ontology by a system based questions/answers [18]: Authors have proposed a semi-automatic approach called SAVANT based on the generation of questions to validate their ontologies. The first step is to automatically generate a list of Boolean questions from the ontology being validated. These questions are submitted to experts in the field who provide an agreement decision (Yes / No) and then an interpretation of these comments made to validate or modify the ontology. The originality of this approach rests on the fact that the interventions are manual and they are carried out only by health professionals.

b) An interactive method for the validation of ontology "OVIM" [19]: An ontology validation method called OVIM "Ontologies Validation by Interactive Method" has been proposed. Authors proposed this method for the structural and semantic validation of ontologies. This method will be based on five stages. They started with the structural validation that has four stages of validation namely; consistency, validation by OOps, validation by request and validation of the choice of the preferential label. In the fifth step they realized the semantic validation by collaborating with actors of the modeled domain.

c) An ontology validation Approach by the experts via a questionnaire [20]: An ontology evaluation and validation approach has been proposed. This approach starts from an ontology to be evaluated and ends with an ontology updated according to the evaluators' recommendations. The proposed approach consists essentially of five steps: In the first step a questionnaire is produced from the components of the ontology. Secondly, results of the survey of the experts will be done. The third step is to analyze and synthesize the results obtained. The update of the questionnaire based on expert feedback as well as the update of the ontology according to the knowledge of the results is realized during the last two stages.

d) A validation approach based on evaluation [21]: This approach essentially consists of verifying the consistency and measuring the impact of the change on the quality of the ontology. It also allows consistency checking and evaluation of the structure and content of the proposed ontology based on well-defined evaluation criteria and metrics.

2) Discussion

Although the validation approach proposed by [18] is a very important approach that allows the validation of concepts, relationships and axiom components of ontology, in fact it has been evaluated experimentally on three ontologies of different methods of construction but this approach presents some lacuna:

- A bad quality of validated ontology is related to two reasons:

1) The absence of a direct interaction between the ontology and experts to validate it [no interface].

2) Wrong time planning of the expert and the reduction of his level of concentration during the answers to the questions.

- The choice of questions is not generic; it also depends on the context of the problem.
- The validation method of [15] like any other method allows the structural and semantic validation. The problem here is that during the semantic validation domain actors verify only the existence of the general semantic domain.
- Another limit of this approach is the fact that the domain experts are not allowed to add, modify or update the used concepts.
- Expert, in this approach are simply domain actor and are not necessarily specialists in the field of ontology engineering.
- The approach proposed in [16], is a very interesting approach but has some limitations:
- It is an approach not updated in the term of the novelties of the version of the OWL language.
- Uses only English for the generation of questionnaires in natural language.

- The questionnaires are generated using non-specialists in the construction of ontology study which reduces the quality of validated ontology.

The study of these different approaches allows us to notice that:

- A total absence of documentation.
- Absence of multi-expert validation [just one expert involved].

- Generally the major approaches make use simply of an evaluation of their ontologies. Effectively this evaluation could not be considered as a validation permitting to exploit really their ontology. It is in this context that we have proposed a so-called incremental validation approach which is mainly characterized by multi-intervention, documentation and incrementation. In the next section, we will describe both the process of building ontology and the proposed validation.

III. Onto-Computer-Project: Methodology of Construction

The proposal of a knowledge capitalization approach is the main goal of our current research study. This approach consists essentially of two processes: a knowledge acquisition and formalization process and a support decision for project management process.

The present paper is only concerned by the first process. It is composed of two phases: The phase of acquisition and the phase of formalization of knowledge. This process is relayed by the proposal of domain ontology wish structures and organizes the great mass of the concepts and knowledge encapsulated in the proposed models.

In a previous work [5] we proposed an ontological construction approach based on the methodolgy "Methontolgy" which leads to a first version of our domain ontology. In this paper, we will:

First detail the description of this approach, by applying carefuly the methodology "Methontolgy" wich has been select in the bases of a comparative study i..e 2.1.2. Finaly, we present a recent version of the ontology [Fig. 6] with our proposed validation approach. The particularity of "Methontolgy" is the possibility of the return on the steps preceding. In what follows, inspired from "Metontology", we will present the stages of the construction of our domain ontology:

- Step1: this step consists in building a glossary of terms containing all the domain knowledge that is useful and potentially usable for the construction of computer domain ontology. This glossary includes concepts, instances, verbs and attributes. To do this step, we have met with domain specialists and experts to talk about computer projects.
- Step2: "classes and class hierarchy": In this step we have built taxonomies of concepts and terms obtained via the grouping, the categorization and the generalization of the different concepts studied [Fig. 1].
- Step 3: "relations between classes" During this stage we have created the relations between classes by determining for each relation the type of relation and the classes to be connected [Fig. 2].
- Step 4: the instantiation of this ontology [individuals + instances] is actually the new case on which our reasoning is based [Fig. 4].
- Step 5: This step provides a detailed description of previously identified relationships, attribute concepts, and constants. We have used the projects of the company and the structure of documents to define some classes and some attributes.
- Step6: "Data properties" It concerns the description of formal rules and axioms relating to the various elements of ontology already known. [Fig. 3].
- Step 7: This steps concerning the detailed description of instances and relations between instances, classes and properties [Fig. 5].

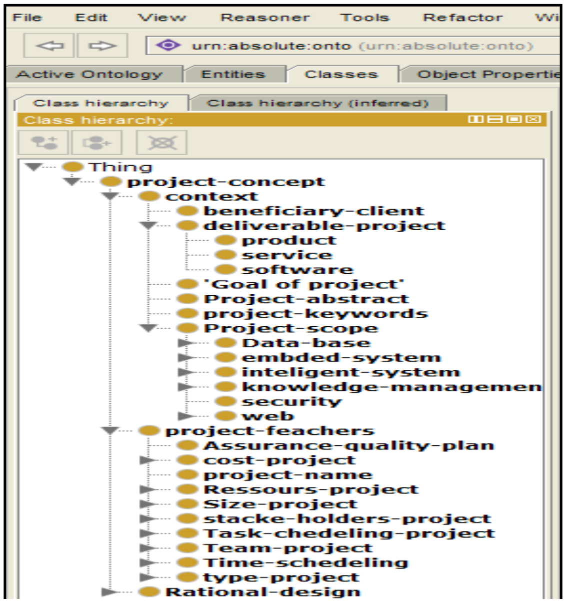


Fig. 1. Class Hierarchy.

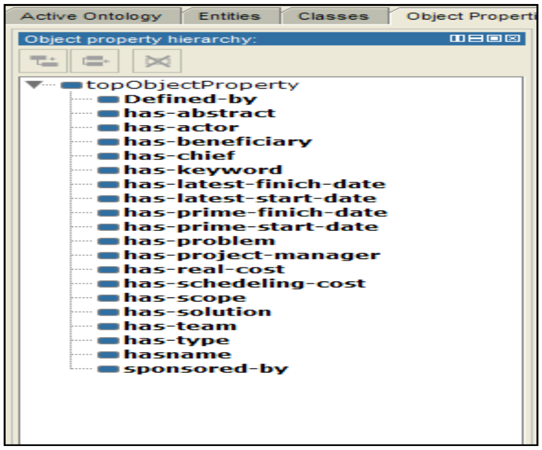
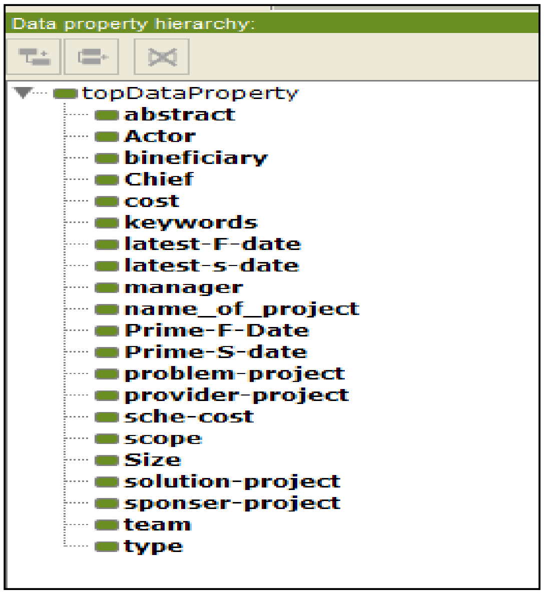


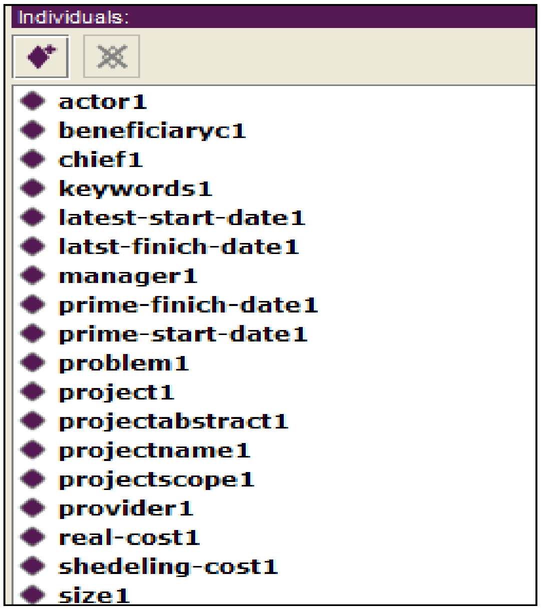
Fig. 2. Object Properties.



Data property hierarchy:

- topDataProperty
  - abstract
  - Actor
  - bineficiary
  - Chief
  - cost
  - keywords
  - latest-F-date
  - latest-s-date
  - manager
  - name_of_project
  - Prime-F-Date
  - Prime-S-date
  - problem-project
  - provider-project
  - sche-cost
  - scope
  - Size
  - solution-project
  - sponser-project
  - team
  - type

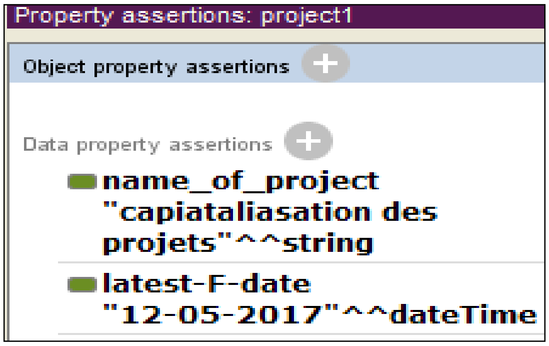
Fig. 3. Data Property of Ontology.



Individuals:

- actor1
- beneficiaryc1
- chief1
- keywords1
- latest-start-date1
- latst-finich-date1
- manager1
- prime-finich-date1
- prime-start-date1
- problem1
- project1
- projectabstract1
- projectname1
- projectscope1
- provider1
- real-cost1
- shedeling-cost1
- size1

Fig. 4. Example of Instantiation.



Property assertions: project1

Object property assertions

Data property assertions

name_of_project "capiatalisation des projets"^^string

latest-F-date "12-05-2017"^^dateTime

Fig. 5. Relation between Instances.

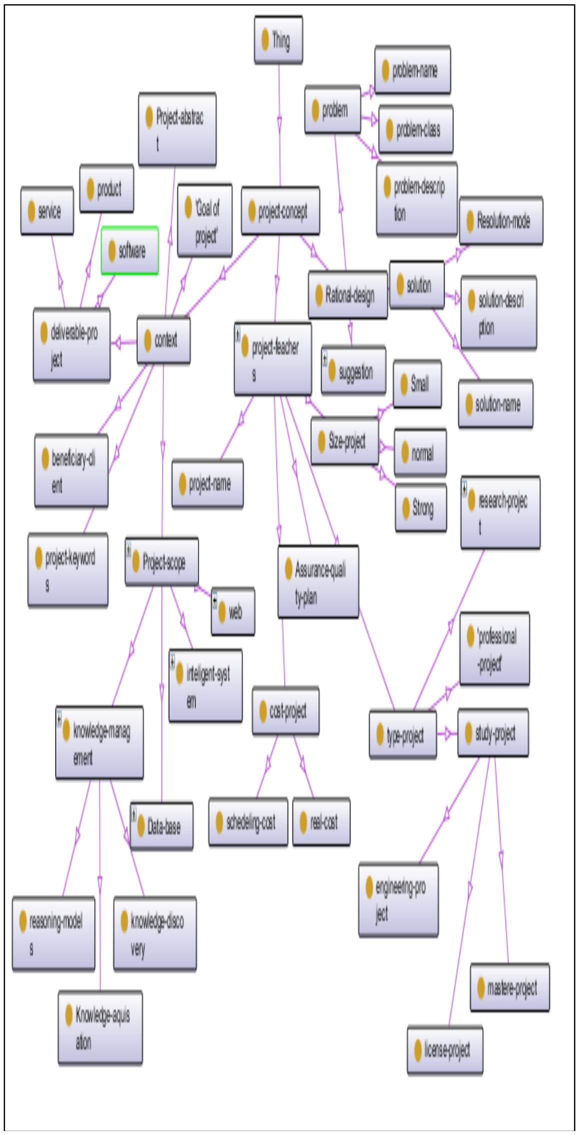


Fig. 6. Onto-Computer-Project: Final Computer Domain Ontology Version.

IV. Incremental and Multi-intervention Validation Approach

Evaluating ontology means checking and validating two aspects: structural aspect and semantic aspects. The validation of the structural aspect of ontology allows verifying the consistency and the coherence of a model to check. In this way, classes and sub-classes are verified according to criteria of consistency and coherence between them and to avoid redundancy.

The validation of the semantic aspects involves communication aspects between actors of different domains of expertise. In this way, we proposed a validation approach based on three criteria:

- The first criterion: the Incremental validation of the ontology: the passage from one validation step to another results in an update [modification, deletion or addition] of the initial ontology.

- The second criterion: the Multi-intervention criteria: This approach is characterized by the intervention of several and different experts. Three experts are involved in the validation process:

✓ The project management expert: He is an expert in the field of project management.

✓ The project computer expert: He is an expert who masters all the concepts of computer projects.

✓ The specialist in ontology engineering: this actor has a good command of all the tools and editors of the ontology.

- The third criterion: our validation approach is respecting the “V cycle”. We inspired by the live cycle of software engineering. Effectively our approach like the V cycle requires a feedback between all the validation phases. Hence, in our validation phases we can return to any expert for revalidation if needed. In contrary of a classic approach which applied semantic and structural validation definitively with no return, we can return at any phase validation to enhance our ontology. The approach that we proposed is essentially composed of six steps [Fig. 7]:

- Step 1: During the first validation step, a descriptive document presented in tabular form containing all the concepts and terms as well as their descriptions constituting our first version of the ontology was be prepared.

- Step 2: In the second step, it is up to us to update our proposal based on the remarks and the assertions given by the computer project expert. This step was considered as a meeting accompanied by discussions. The result of this phase is a second ontology’s version that is ready to be evaluated by "project computer expert". This version is an amelioration of the version 1 at the level of project features [Fig. 8].

- Step 3. During this step, we prepare a second report: a document describing our objectives and orientations. This report is then submitted to a project management expert for evaluation. This second expert could affirm or refute, add or modify the proposal by adding a textual justification. Effectively, in a version 3, this expert proposes to restrict the ontology by adding a new super class named “project context”. This class gives a detailed idea about “project deliverables”, “project abstract” and “project keywords”, etc. [Fig. 9].

- Step 4: After the evaluation done by the project management expert, we have to do at this present step a technical check .this check makes use of a software tool in the way to evaluate the consistency and the coherence of the latest version of our ontology. This mission is assured by a specialist in ontology engineering and results in a version 4.

- Step 5: at this step, the version 4 is sent to the project management expert according to our objective which is essentially to discuss about projects problem solving. Our goal here is to enrich ontology in the way to facilitate problems solve in a new project by exploiting historical projects. This step leads to a new version of ontology labeled version 5. At this effect the expert proposes to add a new sub-class baptized “Rational design” [Fig. 10].

- Step 6: For this validation phase the specialist of ontology engineering chooses to use HERMIT [tool integrated in protégé 4.2] to validate the consistency. This step results in a new version 6.

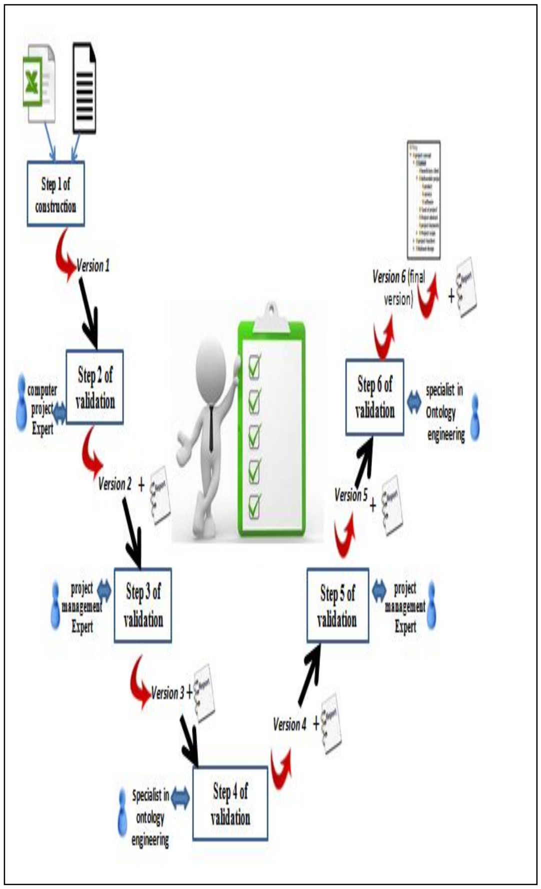
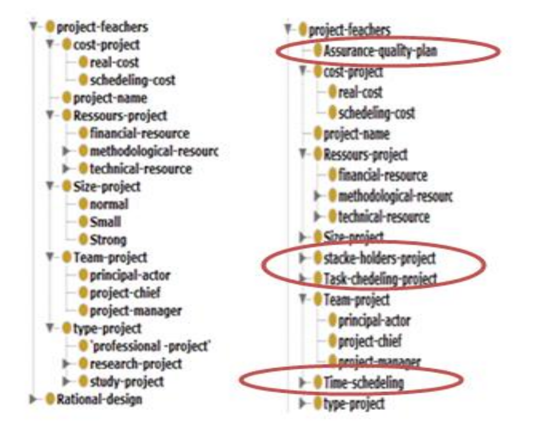


Fig. 7. Incremental Validation Approach.



```
▼ project-feachers
    ▼ cost-project
        real-cost
        schedeling-cost
    project-name
    ▼ Ressours-project
        financial-resource
        ► methodological-resourc
        ► technical-resource
    ▼ Size-project
        normal
        Small
        Strong
    ▼ Team-project
        principal-actor
        project-chief
        project-manager
    ▼ type-project
        'professional -project'
        ► research-project
        ► study-project
► Rational-design
```

```
▼ project-feachers
    Assurance-quality-plan
    ▼ cost-project
        real-cost
        schedeling-cost
    project-name
    ▼ Ressours-project
        financial-resource
        ► methodological-resourc
        ► technical-resource
    ► Size-project
    ► stacke-holders-project
    ► Task-chedeling-project
    ▼ Team-project
        principal-actor
        project-chief
        project-manager
    ► Time-schedeling
    ► type-project
```

Fig. 8. Updating of Project Features.

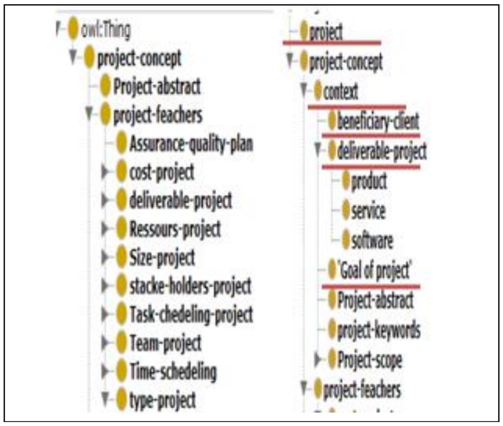


Fig. 9. Adding of Project Context Class.

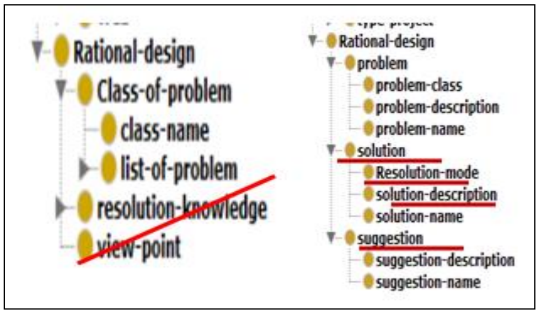


Fig. 10. Rational Design Class Update.

V. Conclusion and Perspectives

Validation ontology plays an important role during the creation and updating of ontology to obtain a final and suitable ontology version. In this paper we have proposed two approaches: an ontology construction approach and a validation approach.

The construction of this domain ontology allowed us to have a complete idea on the concept of projects and specifically computer projects. It also provides support or help for users to acquire new projects that need to be classified. The proposed validation approach is an incremental and a multi-intervention approach that allows a semantic and structural validation of the proposed ontology.

After the validation phase we will validate experimentally this ontology. It is in this context that our near future work will be focused on the experimentation phase. This phase is carried out by building a knowledge base containing a real computer projects forming the basis of the facts and a set of rules forming the basis of the rules. These rules are of two types: classification rules which help to classify the projects and association rules which provide a help to describe in detail a new project.

To do this, we will use the classification data mining techniques and we are going to propose classification and learning algorithms.

References

[1] B.Menaouera , S.Khalissab , B.Abdelbakic , T.Abdelhamidd "Towards a new approach of support innovation guided by knowledge management: Application on FERTIA, 4th International Conference on Leadership, Technology, Innovation and Business Management, 2015.

[2] M.Hemam1 et al, An Ontological Approach for Domain Knowledge Modeling and Querying in Wireless Sensor Networks, The 2nd International Conference on Pattern Analysis and Intelligent Systems, PAIS, 2016.

[3] D.Monticolo et al, An agent-based system to build project memories during engineering projects, Knowledge-Based Systems, 2014.

[4] F. Belkadi, E. Bonjour, M. Camargo, N. Troussier, B. Eynard, A situation model to support awareness in collaborative design, Int. J. Hum.–Comp. Stud. 71 (1), 2013.

[5] H.Raja, M.Lassad, BH.Henda,"Computer-Project-Ontology Construction,Validation and choice of knowledge base",10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 2018.

[6] K. Drame, “Contribution to ontology construction and information retrieval: application to the medical field,” thesis, Université de Bordeaux, 2014.

[7] M.Harzallah,Contributions to Knowledge Engineering: Construction and Validation of Ontology and Semantic Measurements, Habilitation to Direct Research (H.D.R.),2017.

[8] A.Amarir, El.Benlahmer, L.el houssine, "The methods of building ontology from text", Conference Paper, The second day on Information Technologies and Modeling TIM'14, May 2014.

[9] G.Enrico.Caldarola, M.Antonio.Rinaldi, "An Approach to Ontology Integration for Ontology Reuse IEEE 17th International Conference on Information Reuse and Integration, At Pittsburgh, Pennsylvania, 467 Reads. with USA, July 2016.

[10] J.Gherasim, M.Harzallah , G. Berio, P. Kuntz , "Comparative analysis of methodologies and tools automatic ontology construction from textual resources, ∗LABSTICC, UMR 3192 CNRS,2013.

[11] E. Mezghani, E. Exposito, K.Drira, A collaborative Methodology for tacit knowledge management: Application to scientific research, Future Generation Computer Systems, 2015.

[12] Y.Sure,S.Staab,Rudi Studer,On-To-Knowledge Methodology [OTKM], Institute AIFB, University of Karlsruhe Postfach, 76128,Karlsruhe, Germany,http://www.aifb.unikarlsruhe.de,2004.

[13] M.Groninger,M.S.Fox, The Role of Competency Questions in Enterprise Engineering, Department of Industrial Engineering, University of Toronto, 4 Taddle Creek Road, Toronto, Ontario M5S IA4, 2017.

[14] A.Yunianta et al, Methodology for Ontology Development on Data Integration (OntoDI), (IJACSA) International Journal of Advanced Computer Science and Applications, 2019.

[15] H. Qiua,d , G.F. Schneidera,d , T. Kauppinenb , S. Rudolphc S. Steigerd, Reasoning on Human Experiences of Indoor Environments using Semantic Web Technologies,35th International Symposium on Automation and Robotics in Construction, 2018.

[16] L. Trouche, S. Aubin, V. Soulignac, L. Guichard. Construction of a semantic model to organize knowledge dedicated to agroecology. The case of Agro-PEPS / GECO. Agronomy, Environment and Societies, French Association of Agronomy, 2016.

[17] M.Harzallah, Contributions to Knowledge Engineering:Construction and Validation of Ontology and Semantic Measures, Habilitation to Direct Research (H.D.R.),2017.

[18] Ben Abacha et al.,Towards Natural Language Question Generation for the Validation of Ontologies and Mappings, Journal of Biomedical Semantics,2016.

[19] M.Richard, X.Aimé, M.Krebs, J.Charlet, LOVMI: towards an interactive method for the validation of ontologies, INSERM UMRS 1142, LIMICS, F-75006, 2015.

[20] G.Leila,M.aya,D.faiza, generation of a questionnaire from a domain ontology,conference paper, 2017.

[21] S.tartir,S.Amit,IA.Young,Ontological Evaluation and Validation, from book Theory and Applications of Ontology: Computer Applications [pp.115-130], 2010.

Comparison of Item Difficulty Estimates in a Basic Statistics Test using ltm and CTT Software Packages in R

Jonald L. Pimentel[1*]
Department of Mathematics and Statistics
University of Southern Mindanao
Kabacan, Cotabato, Philippines

Marah Luriely A. Villaruz[2]
Ascend E-Commerce Phils., Inc.
Pasig City
Philippines

***Abstract*—Two free computer software packages "ltm" and "CTT" in the R software environment were tested to demonstrate its usefulness in an item test analysis. The calibration of the item difficulty parameters given the binary responses of two hundred five examinees for the fifteen items multiple choice test were analyzed using the Classical Test Theory (CTT) and Item Response Theory (IRT) methodologies. The software latent trait model "ltm" employed the IRT framework while the software classical test theory functions "CTT" operated under CTT. The IRT Rasch model was used to model the responses of the examinees. The conditional maximum likelihood estimation method was used to estimate the item difficulty parameters for all the items. On the other hand, all the item difficulty indices using the "CTT" software were also calculated. Both the statistical analyses of this study were done in the R software. Results showed that among the fifteen items, the estimates of their item difficulty parameters differed mostly on their values between the two methods. In an IRT framework, items showed extreme difficulty or easy cases as compared to CTT. However, when the estimated values were categorized into intervals and labelled according to its verbal difficulty description, both methodologies showed some similarities in their item difficulties.**

***Keywords*—Classical test theory; indices; item calibration; item difficulty; item response theory; R software**

I. Introduction

In the field of education particularly in test and measurement, it is important that any method that uses technology should be upgraded from time to time. This technology that performs computing and analysis requires speed and precision especially if the data is huge. Hand computation seems to be tedious and possible but it will take a longtime. In the case of a test or item test analyses, it is important that the item calibration for the estimates of its item parameters is accurate, fast and reliable. Statistical software packages that perform these calculations are available, either purchased commercially or as a free software in the internet.

Test item analysis is very important especially in the test construction. First, test can be classified with its degree of difficulty and second, for item banking that is, the calibrated items are stored traditionally in a box or electronically in a database. These items were given labels for its corresponding levels or index of difficulty of which it can be retrieved anytime for test construction. This method is useful for test makers in the composition of test items and the determination of the difficulty or easiness of the test instrument.

The primary objective of this paper is to demonstrate the usefulness of the two computer software programs, the latent trait model "ltm" [1] and the classical test theory functions "CTT" [2] in the R software environment in the calibration of item difficulty parameter estimates/indices for a multiple-choice test. Two methodologies, the item response theory and the classical test theory will be used. Specifically, this study will employ the Rasch model [3], an IRT probabilistic model which is part of the logistic model family, to model the responses of all the examinees for all the items. Estimation for all the item difficulty parameters will be carried using the conditional maximum likelihood estimation [4]. The calculated item difficulty estimates will be compared to the calculated difficulty indices of the same test examination that uses the scores of the examinees under the classical test theory methodology. One point of interest in this study is the comparison of the verbal description of the items in terms of its difficulty labels. Here we will know whether each item estimates are comparable for both methodologies or they both possess extreme differences.

II. Background of the Study and Related Studies

A. Item Calibration

In the calibration of item parameters, specifically the difficulty indices β of an examination test say in the case of a multiple-choice test in which the resulting data is a matrix of binary responses of the number of examinees who took the examination and the number of items being answered, Two methodologies are available at present in the literatures to handle such calibration. These methods are the Classical Test Theory (CTT) which is based on prediction of outcomes on a test that is, in particular an examinee's observed score which is composed of a true score and an error score and the Item Response Theory (IRT) which is based on a response probabilistic modeling [4]. CTT usually do the estimation of the reliability of a test and the item difficulty indices which comes from the score of the examinees. In practice, these indices are also known as the p-value and is valued from 0.0 to 1.0 for each item and it is based on the proportion of all the examinees who got the correct answers over the total

[*] Corresponding Author

examinees. The higher the proportion of getting correct answers, the easier is the item. CTT however, has many limitations as cited by [5]. On the other hand, test makers are also adopting model based IRT because it is powerful and can provide a framework for evaluating how good assessment do its job and how good its item do its job. For calculating item difficulty for example in a multiple-choice test, IRT traditionally applied based on a large number of historical correct or incorrect information gathered from the test [6] and in turn applies probabilistic models as mention by [4]. See also [7], [8], [9] and [10] for more discussion about these item response theory modeling.

B. Available Statistical Software Programs

Statistical software packages are presently available for calibrating item parameters in which CTT and IRT models are used. These includes powerful commercial software such SAS [11],_STATA [12], SPSS [13], M plus [14], the BILOG-MG software [15] and ConQuest software [16] for fitting item response latent regression models and many more. However, there are some commercial software packages that are not easy to learn as well, hence it is must to do an extensive training if you want learn it because some of the software corresponding documentations are difficult to comprehend and sometimes have program failures and limitation which can be frustrating.

There is also a software package developed by The National Institute for Educational Measurement of the Netherlands (CITO) called OPLM [17] which is free and can be obtained by request. Starting in the year 2000, a quite number of new IRT packages uploaded as a library were developed in the open source in R software environment [18]. These includes computer software programs called the latent trait model (ltm) which was intended for unidimensional item response theory [1] as mentioned earlier, the extended Rasch models called eRm [19], the software called mirt which is intended for multidimensional IRT [20], and the software called mlirt which is intended for multilevel and Bayesian estimation [21]. Also, a software in R that uses Bayesian methods is also available called R2WinBuGs [22]. Lastly the software called "CTT" is intended for the estimation of items parameters under the classical test theory methodology [2].

III. Material and Methods

A. The Dataset

The data used in this study were the responses of two hundred five (205) students who responded to a fifteen (15) items multiple choice test. The test was just part of the Basic Statistics Preliminary Examination of the Mathematics and Statistics Department of the College of Science and Mathematics, Mindanao State University –Iligan Institute of Technology during the second semester school year 2014-2015 [23]. The test questionnaire was made by the authors and was validated for its content. The responses of these students were tabulated in a 205 by 15 matrix of 1's, when student got correct answer to the given item and 0's, when student got a wrong answer to the given item (see Table I for the illustration).The data then was stored as a text file having file name. The data was processed using a personal computer.

TABLE. I. Illustrating the Data Matrix

examinee	Item 1	Item2	…	Item 15
1	1	0	.	0
2	0	0	.	1
.	.	.	.	.
.	.	.	.	.
205	1	0	.	1

B. Item Response Theory Models for Binary Response

In an IRT framework, one can specify the components affecting the probability that an examinee will respond in a particular way to a particular test item. We can choose a particular measurement model that will relate the responses of the examinees and the qualities of the items. In this study the Rasch model was the model used to obtain the estimates of the item difficulty applying the conditional maximum likelihood estimation method. This model is one of the simplest item response theory models [24]. In general, the model is characterized as a two parameter models with the ability parameter of the examinee and the other parameter is the characteristics of the item which is the difficulty parameter [25]. The model is given by.

$$P\left(X_{nj}=1\middle|\theta_n,\beta_j\right)=\frac{e^{(\theta_n-\beta_j)}}{1+e^{(\theta_n-\beta_j)}}$$

where X_{nj} refers to a response made by an examinee n to an item j, θ_n refer to the trait level or ability of an examinee n; and β_j refers to the difficulty characteristics of the item j and it may take values between -3 (Very easy) to 3 (very difficult). The expression, $P\left(X_{nj}=1\middle|\theta_n,\beta_j\right)$ is the chance or the likelihood that an examinee n will give an answer to an item correctly conditional to his ability (θ_n) and the difficulty of the item (β_j). It is common in IRT that all measurements in the ability and difficulty before being subjected to estimation under the Rasch model are transformed into standard normal so normal measurements will be used. In the Rasch modeling, an examinee's answer in a form of a dichotomous response (that is, in our data, 1 refers to an examinee who got a correct response to an item while the entry 0, means that the examinee got a wrong answer) can be explained by the examinee's ability and the difficulty characteristic of the item. Now, in order for the model to be generalized, assumptions are considered that will make the model hold. Please see [26] and [27] for more explanations. The majority of applications of item response theory models usually to categorical data as known in [3], [7], [8], [9] and [10] but they were also being applied to data where there are continuous responses. These can be seen in the literatures [28] and [29].

C. Information Characteristics Curve (ICC) under the Item Reponse Theory

The Information Characteristics Curve (ICC) represents the item response function (IRF) which is the likelihood or chance of getting a positive response to each item which is represents the function of the proficiency or ability θ of the examinees. One can represent in the same graph the observed and the expected ICCs to get the fit of each item [30].

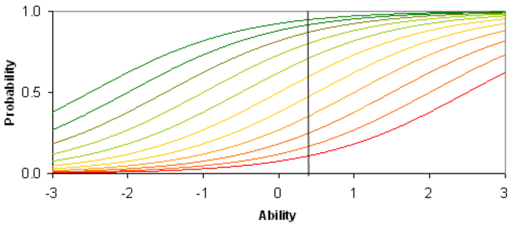


Fig. 1. Information Characteristics Curves (ICCs).

Fig. 1 illustrates ICCs for a number of items. ICCs highlight the change in the chances or likelihood of a successful response for an examinee with its ability location at the vertical line. The examinee will likely respond correctly to the easiest items (with locations to the left and higher curves) and unlikely to respond correctly to difficult items (locations to the right and lower curves) that is, the x-axis is the theoretical ability or proficiency level, ranging from -3 to +3. This graph only represents theoretical modeling rather than empirical data. To be specific, there may not be examinees that can reach a proficiency or ability θ level of +3 or fail so miserably as to be in the -3 group. Nonetheless, to study the characteristics of an item, we are interested in knowing, given a person whose θ is +3, what the probability of giving the rights answer to an item. The ICC indicates that when θ is zero, the examinee is on an average ability or proficiency hence, the chances of the examinee of answering the item correctly is approximately 0.5 or 50%. When the ability level θ is -3, the probability is almost zero to correctly get the item. When θ is +3, the probability to correctly answer the item increases to 0.99 or 99%.

D. *Maximum Likelihood Estimation for the Item Difficulty with the Rasch Model*

In order to calculate the values of the estimates of the item difficulty parameters, the computer software program “ltm” which means Latent Trait Models under the IRT framework [1] will be used in the calibration of item difficulty estimates and is based on the environment of the software R. Further, the “ltm” adapted both the conditional maximum likelihood (CML) and marginal maximum likelihood (MML) estimation methods that handles the calculation in estimating item parameters and person parameters mathematically. In our study we employed the conditional maximum likelihood to calculate the item difficulty parameter. For more details, please see [6], [27], [31] and [32]. Although, maximum likelihood methods are the common estimation methods for years in the calibration of examinee’s proficiency or ability and item parameters particularly the discrimination, difficulty and the guessing parameters another alternative estimation method emerged. The development of the Bayesian framework as an alternative but very powerful sampling-based estimation techniques have encouraged the application of Bayesian methods. The Markov Chain Monte Carlo (MCMC) methods, such as Gibbs sampling and Metropolis-Hastings (M-H), were used to simultaneously estimate all model parameters. An MCMC implementation are introduced for the sampling of all model parameters that combines various advantages of different MCMC schemes for sampling IRT parameters [33]. For example [34] estimated the ability and item parameters of the IRT model for the observed data using MCMC.

The Bayesian inference requires the computation of the posterior distribution for a collection of random variables (parameters or unknown observables). At present, numerous simulation-based methods emerged. On sampling by [35] and [36], Other Bayesian works see [37] and [38]. Statistical software packages like WinBUGS (Bayesian inference Using Gibbs Sampling) [39] is a popular software for analyzing complex statistical models using MCMC methods.

E. *Classical Test Theory*

The classical test theory (CTT) is a theory of measurement error. The classical test theory has an assumption that each examinee’s actual observed score X is the sum of the examinee’s true score T and the error score E that is X=T+E. The key concepts of this theory involved the determination of test’s reliability and validity for which test can be assessed mathematically [40]. The study and application for the classical test theory has been continuing which can be seen in the literatures [41]. Further, major applications of this theory are also on the test and item analysis and observed score equating. An article published in [42] looks for working on the classical test theory in combination with the concept of the item response theory. Their paper, emphasized that since the classical test theory was built in the assumption of exchangeability and the item response theory was based on conditional independence then they concluded that item response theory can be considered as an extension of the classical test theory where the concepts for both theories are related with each other. What is interesting in their work is the capability of IRT to provide the classical test theory statistical values where it can provide.

In our study, the software package “CTT” will be used to calculate the item difficulty indices of the test which is based on the proportion of the total number of examinees who got a correct answer on the given item and the total number of examinees. The closer the value of the index of difficulty of the given item to 1, the easier is the item and the closer it is to 0 the item will be very difficult. An index of 0.5 means that the item is average in its difficulty. We will also categorize the different intervals so that item difficulty indices can be given a verbal description. Moreover, in this study for reasons of simplicity and completeness of the estimation of item difficulty parameter for each of the 15 items, we assume that the item response theory’s Rasch model fits the data, that is in every responses of examinees on each item fits in the model. Although, we will check the goodness of fit of the items to the model by statistical means as done by [43] in the process. For the sake of comparison, we do not discard items in the estimation that do not fit the given item response model, that is we need the complete 15 item difficulty estimates so we can compare it to the values in the classical test theory.

IV. Results and Discussions

This section will present three results. First is the presentation and discussions of the information characteristic curve (ICCs) of the fifteen items under the item response theory methods. Second and third is the simultaneous tabular presentation of the calculations and analysis of the results for the estimation of the item difficulty under the item response theory (IRT) and the classical test theory (CTT) methods.

A. *The Information Characteristics Curves of the Items*

Fig. 2 are the information characteristic curves of the fifteen items, the ICCs of the fifteen items above can be converted into an Item Characteristic Curves (ICC) which are graphical functions that shows the examinees proficiency or ability as a function of the likelihood or chances of answering correctly the item. We can see through the curve that most of the items (9 out of 15 items) are difficult because higher abilities are needed to get higher probabilities of getting correct answer. We can see in the figure that items 8 and 12 are the very difficult items. To further support these observations, we will calculate mathematically using the Rasch model under the item response theory methods the values of the estimates of the difficulty of the fifteen items. Then we will incorporate the results of fifteen difficulty indices of the items under the classical test theory.

B. *Fit of the Item to the Rasch Model*

Checking the fit of the data to the Rasch model, the results show that some items are a “misfit”, a terminology in modeling for those items that do not fit the model. As we mentioned above those items that do not fit in the Rasch model are supposedly discarded but for the purpose of comparison with difficulty indices under the classical test theory we will retain it. Table II shows those items that fit and also do not fit the Rasch model. To test the fit of the data responses of the item to the model, we use the Chi-square test. If the p-value of that item is less than 0.05 or 5%, we do not reject the hypothesis that the item fits the model. Based on the results in Table II, the following items fit the model in particular items 2, 3, 7, 10, 11, and 12. On the other hand, items 1, 4, 5, 6, 8, 9, 13, 14, and 15 in the test did not fit the model. Items that misfit the Rasch model means that the Rasch model is not a good model for these items, hence for model fitting purposes, another type of item response theory should be considered as a recommendation.

C. *Comparison of Item Difficulty Estimates Between IRT and CTT*

The item difficulty estimates of the data using the IRT and CTT methodology are presented in Table III.

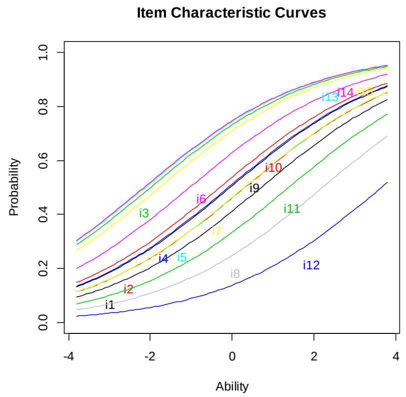


Fig. 2. ICCs of the 15 items using IRT Rasch Model.

TABLE. II. FITTING THE DATA USING RM (MODEL FIT)

Items	Test	
	X^2	p-value
1	18.49	0.01*
2	5.66	0.46
3	11.51	0.07
4	14.41	0.03*
5	22.23	<0.01*
6	24.78	<0.01*
7	9.97	0.13
8	23.08	<0.01*
9	15.16	0.02*
10	11.84	0.07
11	7.70	0.26
12	7.49	0.28
13	14.01	0.03*
14	34.71	<0.01*
15	22.62	<0.01*

Legend: * Significant at 5% (using chi-square test)

TABLE. III. ITEM DIFFICULTY ESTIMATES/ INDICES

Items	β (IRT)	description[a]	β (CTT)	description[b]
1	0.44	D	0.40	A
2	0.21	D	0.45	A
3	-1.01	VE	0.71	E
4	0.01	A	0.50	A
5	0.23	D	0.45	A
6	-0.59	E	0.62	A
7	0.21	D	0.45	A
8	1.23	VD	0.25	D
9	0.03	A	0.49	A
10	-0.10	E	0.52	A
11	0.75	D	0.34	D
12	1.91	VD	0.16	D
13	-1.09	VE	0.72	E
14	-1.20	VE	0.74	E
15	-0.96	VE	0.70	E

Legend: β=item difficulty values IRT ∈ [-3,3], CTT ∈ [0,1]
[a]Description: VE = Very Easy, E = Easy, A = Average,
D = Difficult, VD = Very Difficult
[b]Description: VD=0-0.125, D=0.126-0.375, A=0.376-0.625
E=0.626-0.875, VE=0.876-1.0

Includes corresponding verbal descriptions of all the items. Discussing the item difficulty estimates of the fifteen items that was included in the test under the IRT framework that employed the Rasch model, results showed that the level of difficulties, the test in particular items 1, 2, 5, 7, and 11 are difficult items because they are above 0. Note that an item whose difficulty is zero is considered an average item. Items 8 and 12 can be considered very difficult items because they are almost near at the upper right extreme.

Further, items 4 and 9 are items on an average difficulty. On the other hand, items 6 and 10 can be considered easy items while items 3, 13, 14, and 15 are very easy items because they are in extreme left near the value -3 considered the easiest item. In the case with CTT methods, results of the analysis show that items 8,11 and 12 are difficult items and items 1, 2, 4, 5, 6, 9 and 10 are items with average difficulty while the rest of the items, items 3, 13, 14 and 15 are easy items.

For the point of comparisons in accordance to the item difficulty estimates calculated under the two methodologies in particular, the items verbal descriptions, with regards to classical test theory (CTT), results revealed that there were no very difficult items in the test. Examination and further there were also no very easy items. This can be explained maybe due to the choice of the categorized interval from 0 to 1. Three items out of the total fifteen items namely items 4, 9 and 11 showed similar descriptions in their item difficulty for both item response (IRT) and classical test theory (CTT) methodologies. However, there are also some items that do not have the same common description (about 9 out of 15 items or 60% of the items). These items are items 1, 2, 3, 5, 6, 7, 8, 10 and 12. These results are expected since the two methodologies have different assumptions in their formulations.

As we mentioned above, the assignment of the degree of difficulty depends on the kind of interval that was made. As we observed, some intervals are narrow and some are wide. In the case of item 1, it is difficult under the IRT formulation but is an average item in CTT. It is also the same result with items 2, 5, 7 while item 3 is very easy in IRT but is an easy item in CTT. Items 6 and 10 are both easy in IRT but were average items in CTT while items 8 and 12 are very difficult items in IRT but only difficult items in CTT and lastly items 13,14,15 are very easy items in IRT but are easy items in CTT. A study by [44] compared CTT and IRT for the examinee change assessment. According to them a lot of investigators were eager to know of how IRT can be used in greater advantage as compared to CTT in change assessment but available results showed that they did not differ when compared based the examinee change assessment. However, when compared in term of their type 1 errors and detection percentages, their results showed that IRT is better than CTT in the examinee ‘s change detection with the condition that the test must consists twenty (20) items or more. For shorter tests, however they further mentioned that CTT has the advantage of correctly knowing change in the examinees. In our study, however there was also some variations in the results between IRT and CTT among the item difficulties when they were compared but the objective of this study was achieved. The two free computer software programs the “ltm” and “CTT” were very useful in doing the statistical analysis using the R software for the item test calibrations.

V. Conclusions and Recommendations

This paper demonstrated the usefulness of the free computer software programs, the “ltm” and “CTT” in the R software environment for the calibration of the item difficulty parameter estimates/indices of the multiple-choice test examination using both the item response theory (IRT) and classical test theory (CTT) methodologies. We also demonstrated the usefulness of the Rasch model, an IRT probabilistic logistic model used to estimate the values of the item difficulty parameters of the test examination which were compared to the estimated values under the CTT method. Further, we also demonstrated that it was possible to plot the item characteristics of different items, so the proficiency or ability of the examinee can be estimated so that we will be able to know the higher chance of getting the item correctly. The Item characteristic curves (ICC) also gave us a glimpse of the difficulty characteristic of the item. The study also found some differences and similarities in the interpretation with the labeling of the item difficulty in the form of a verbal description for the items. The study concluded that these differences are due to the assumptions of the different methods in the item analysis.

The study further concluded that for the possibility of convenience for teachers in all levels and test constructors, they can do an item analysis for their test electronically using either ltm or CTT software packages in R for free in which first, they can do item calibration and assigned description for the item level of difficulty or indices and second, for the purpose of item banking especially in the test construction where items are stored in a database and labeled with their corresponding item level difficulty or indices.

This study further recommends that other item characteristics namely, the item discrimination and the item guessing parameters shall also be investigated to complete the test item analysis using both the classical test theory functions and other appropriate item response theory models that involves item calibration for discrimination and guessing.

Acknowledgment

Our heartfelt thanks to all the students who participated and have tried their best to answer the different items in the test. The first author also expressed his gratitude to his employer, the University of Southern Mindanao and his current President Dr. Francisco Gil N. Garcia for the encouragement and support.

References

[1] D. Rizopoulos, “ltm: An R Package for latent variable modeling and item response analysis,” Journal of Statistical Software, vol.17, no.5, 2006.

[2] J.T. Willse, Classical Test Theory Functions, 2018. https://cran.rproject.org/web/packages/CTT.

[3] G. Rasch, Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960.

[4] J.L.Pimentel, Item Response Theory Modeling with Nonignorable Missing Data. University of Twente, The Netherlands ISBN:90-365-2295-1, 2005.

[5] R.K. Hambleton, H. Swaminathan and H.J. Rogers, Fundamentals of item response theory, Newbury Park, CA: Sage, 1991.

[6] F.B. Baker, Item response theory: Parameter estimation techniques. New York, NJ: Dekker, 1992.

[7] F. Samejima, “Estimation of latent proficiency using a pattern of graded scores,” Psychometrika, Monograph Supplement, no. 17, 1969.

[8] R.D. Bock, "Estimating item parameters and latent ability when responses are scored in two or more nominal categories," Psychometrika, vol. 37, pp. 29 – 51, 1972.

[9] F.M.Lord, Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

[10] G.N.Masters, "A Rasch model for partial credit scoring," Psychometrika, Vol. 47, pp.149 – 174,1982.

[11] R.Codey and J.K. Smith, Test Scoring and Analysis Using SAS. SAS Institute Inc., Cary, North Carolina, USA, 2014.

[12] J.S.Yang, X. Zheng, "Item Response Data analysis using Stata Item Theory Package," Journal of Educational and Behavioral Statistics, vol. 43, no. 1, pp. 116–129, 2018. Available; DOI: 10.3102/1076998617749186 © 2017AERA.

[13] P.J. Pascale and J.S. Pascale, "Item Analysis Using the Statistical Package for the Social Sciences (SPSS),"Educational and Psychological Measurement, vol.40 no.1, pp.163-164,1980.

[14] L.K. Muthen and B.O. Muthen, MPLUS: The comprehensive modeling program for applied researcher, users guide. Los Angeles, CA: Muthen & Muthen, 1998.

[15] M.F. Zimowski, E. Muraki, R.J. Mislevy and R.D. Bock, Bilog-MG. Lincolnwood, IL, Scientic Software International Inc.,2002.

[16] M.L.Wu, R.J. Adams, and M.R. Wilson. ConQuest: Multi-Aspect Test Software, Camberwell: Australian Council for Educational Research, 1997.

[17] N.D. Verhelst, C.A.W. Glas and H.H.F.M.Verstralen, OPLM: computer program and manual, Arnhem: Cito, the National Institute for Educational Measurement, the Netherlands, 1995.

[18] R Core Team, R: A language and environment for statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2013.

[19] P. Mair and R. Hatzinger, "Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R," Journal of Statistical Software, vol. 20 no.9, 2007.

[20] R.P. Chalmers, "mirt: A multidimensional item response theory package for the R environment," Journal of Statistical Software, vol. 48 (6), 2012.

[21] G.J.A Fox, "Multilevel IRT modeling in practice with the package MLIRT,"Journal of statistical software, vol. 20, no.5, 2007.

[22] S. Sturtz, U. Ligges and A. Gelman, A."R2WinBUGS: A Package for Running WinBUGS from R," Journal of Statistical Software, vol.12, no.3 , pp.1–16, 2005.

[23] M.L.A.Villaruz, Test Item Calibration for Multiple Choice Test Using IRT, Unpublished Undergraduate Thesis. MSU-IIT, Iligan City, Philippines, 2015.

[24] I.W. Molenaar, "Estimation of item parameters", In G.H. Fischer,and I.W. Molenaar (Eds.), Rasch models: foundations, recent developments and applications, New York, NJ: Springer, 1995.

[25] M.G.H. Jansen, "A Model for the Latent Traits in Rasch's Speed Tests," Applied Psychological Measurement, vol. 27, no. 2. pp.128-151, 2003. DOI: 10.1177/0146621602250536.

[26] G.H. Fischer, "Derivations of the Rasch model". In G.H. Fischer & I.W. Molenaar (Eds.), Rasch models: foundations, recent developments and applications, pp.39-52, New York, NJ: Springer,1995.

[27] G.H.Fischer and I.W. Molenaar, Rasch models. Their foundation, recent developments and applications. New York, NJ: Springer. pp.44-49, pp. 219-224, 1995.

[28] G.J. Mellenbergh, "A Unidimensional Latent Trait Model for Continuous Item Responses," Multivariate Behavioral Research, vol. 2, no.3, pp. 223-236, 1994.

[29] A. Skrondal and S.Rabe-Hesketh, Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models, Chapman and Hall/CRC,2004.

[30] F.B.Baker, The Basics of Item Response Theory. Second Edition. 2001.

[31] S.E. Embretson and S.P. Reise,. Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum. pp.210-218, 2000.

[32] R.K. Hambleton and H. Swaminatan, Item response theory: Principles and applications (2nd edition). Boston MA: Kluwer Academic Publishers, 1985.

[33] J. Fox, J.Pimentel and C.Glas, "Fixed Effects IRT Model", Behaviormetrika 33, pp. 27-42, 2006.

[34] A.E. Gelfand and A.F.M. Smith, "Sampling-based approaches to calculating marginal densities," Journal of the American Statistical Association, vol. 85, pp. 398-409, 1990.

[35] M..H. Chen, Q.M.Shao and J.G. Ibrahim, Monte Carlo methods in Bayesian computation. New-York: Springer-Verlag, 2000.

[36] B.D. Ripley, Stochastic simulation, New York: Wiley,1987.

[37] C.P. Robert and G. Casella, Monte Carlo statistical methods, New York, NY: Springer,1999.

[38] A. Gelman, J.B. Carlin, H.S. Stern and D.B.Rubin, Bayesian data analysis, London: Chapman and Hall, 1995.

[39] D.Spiegelhalter, A. Thomas, N.Best, and D. Lunn, WinBUGS User Manual, MRC Biostatistics Unit, Cambridge, 2003.

[40] M.R. Novick "The axioms and principal results of classical test theory," Journal of Mathematical Psychology, vol. ,3, no.1, pp. 1-18,1966.

[41] R. Traub, "Classical Test Theory in Historical Perspective," Educational Measurement: Issues and Practice, vol. 16 no.4, pp.8–14, 1997. doi:10.1111/j.1745-3992.1997. tb 00603.x.

[42] T.M. Bechger,G. Maris, H.F.M Verstralen, and A.A. Beguin, "Using Classical Test Theory in combination with Item Response Theory", Applied Psychological Measurement, vol. 27 No.5. pp. 319-334, 2003. doi: 10.1177/0146621603257518.

[43] R.M. Smith, "Theory and practice of fit". Rasch Measurement Transactions, vol.3 (4) pp. 78, 1990.

[44] R. Jabrayilov, W.H.M. Emons, and K. Sijtsma, "Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment,"Applied Psychological Measurement vol.40,no.8,pp.559-572, 2016.

Mosques Smart Domes System using Machine Learning Algorithms

Mohammad Awis Al Lababede[1]
Computer Science Department
Mutah University
Al Karak, Jordan

Anas H. Blasi[2]
Computer Information Systems Department
Mutah University
Al Karak, Jordan

Mohammed A. Alsuwaiket[3]
Computer Science and Engineering Technology Department
Hafar Batin University
Hafar Batin, Saudi Arabia

***Abstract*—Millions of mosques around the world are suffering some problems such as ventilation and difficulty getting rid of bacteria, especially in rush hours where congestion in mosques leads to air pollution and spread of bacteria, in addition to unpleasant odors and to a state of discomfort during the pray times, where in most mosques there are no enough windows to ventilate the mosque well. This paper aims to solve these problems by building a model of smart mosques' domes using weather features and outside temperatures. Machine learning algorithms such as k-Nearest Neighbors (k-NN) and Decision Tree (DT) were applied to predict the state of the domes (open or close). The experiments of this paper were applied on Prophet's mosque in Saudi Arabia, which basically contains twenty-seven manually moving domes. Both machine learning algorithms were tested and evaluated using different evaluation methods. After comparing the results for both algorithms, DT algorithm was achieved higher accuracy 98% comparing with 95% accuracy for k-NN algorithm. Finally, the results of this study were promising and will be helpful for all mosques to use our proposed model for controlling domes automatically.**

***Keywords*—Decision tree; k-nearest neighbors; smart domes; weather prediction; machine learning**

I. Introduction

Islam is the second largest religion after Christianity in the world, according to a study conducted in 2015, Islam have 1.9 million followers in the world, representing 24.8% of the world's population [1]. In addition, there are 3.6 million mosques around the world [2]. In fact, mosques have a major problem of a good ventilation due to the crowd since there are many worshipers inside the mosque, and as the windows are not enough for fully ventilation, in addition to the problem of the presence of bacteria and moisture on the mosques' carpets.

In general, the mosques need to keep up with technology and evolution, even if moving domes are exist in some mosques, they are inefficient and there are many problems that prevent them to stay open for a long time, such as weather change and weather conditions, where the weather suddenly turns from clear to dusty, rain, hurricane problems, sandstorms and strong sunlight that disturb worshipers, also the temperatures are high or low outside, so it is difficult to control these domes manually and there are many factors that make the decision hard to open or close the domes.

This paper introduces solutions to previous problems by creating smart moving domes considering weather prediction. However, weather can be predicted by studying satellite data or the behavior of animals affected by weather changes and weather maps [3]. To make this research applicable and reliable, the experiments in this paper will be applied on Prophet's Mosque in Saudi Arabia, which basically contains twenty-seven manually moving domes. Where a Prophet's Mosque is one of the largest mosques in the world and the second holiest Islamic site which was built by the messenger of Allah; Mohammad -peace be upon him- in 1 Hijra, and then was expanded several times throughout the history by the princes of Islamic countries for each period and the largest expansion was during the reign of Saudi Arabia in 1994. The domes were built using silver, granite, gold and marble. Moreover, Prophet's Mosque is expanded to suit 707.000 worshipers, and it is visited by more than 278,000 Muslim every hour from all over the world [4].

Some requirements have been proposed for the domes system to make them move and some required materials should be considered such as:

- Steel rails*:* should be made to handle the friction caused by the moving domes, where the domes are placed on these rails for opening and closing.
- Arduino controller: a simple Arduino device is needed to send (on, off) signals, which gets information from a processing device that forecasts weather and dome state.
- Processing device*:* a computer or processing cloud used to implement the weather prediction algorithm and the dome state.
- Machine learning algorithms: to predict the weather and the domes state, a model should be built using machine learning algorithms such as Decision Tree (DT) and k-Nearest Neighbors (k-NN), where weather factors such as visibility, temperature, humidity, strong wind, clock and barometer will be used to make the decision to open or close the domes.
- Rainfall sensor: to minimize the error rate of the proposed model, it will be used to detect rainfall at the real time and force the domes to close.

In this paper, a model of smart domes will be built using Machine Learning algorithms such as Decision Tree (DT) and k-Nearest Neighbors (k-NN), considering some weather factors like minimum temperature, maximum temperature, wind speed direction, humidity, and average dew point [5]. The main aim of this paper is to solve the problem of ventilation of the mosques by controlling the domes automatically using ML algorithms instead of controlling them manually.

The paper is organized as follows: Section II reviews the related work of weather forecasting using different ML techniques, section III describes the process followed to prepare the data including data understanding, selecting, transforming, and model building. Section IV describes the interpretation and evaluation of the results. Finally, section V discusses the conclusions and draws the future work.

II. Related Work

In this section, some related work will be presented and reviewed to show how others have applied machine learning techniques such as Decision tree (DT), k-Nearest Neighbors (k-NN), Artificial Neural Networks (ANN), Linear Regression (LR), Naive Bayes (NB) and other data mining techniques [5] to predict the weather changes.

According to the studies [6, 7] Artificial Neural Network (ANN) and Deep learning are gaining much popularity due to its supremacy in terms of accuracy when trained with huge amount of data. Deep learning outshines several other artificial intelligence techniques when there is lack of domain expertise in feature engineering, or when it comes to complex problems such as optimization, image classification, natural language processing, and speech recognition.

Some other studies concern about nature events such as [8], authors studied the micro seismic events were detected and classified through combining ML algorithms such as back up vector machine, MLP, NN, C4.5 decision tree and k-NN in the form of boost learning. The procedure of experiment was in a way that less important and important seismic events caused by weight falling from various heights and different distances of far, middle and near recorded by laboratory devices and sensors. Then, the data were pre-processed and classified. The classification was based on the level of height, distance and considered sensors. The precision and accuracy were significantly improved by this strategy. After simulation of their proposed method, it was observed that the precision of proposed boost method was improved up to 6.1% compare to the other methods. The error rate improved up to 0.82% and the recalling and accuracy of detection and classification to the best answer were also improved in the proposed method up to 2.31% and 6.34%, respectively.

According to [9], the authors have built a hybrid system between a multilayer perceptron (MLP) and Radial basis function (RBF) to enhance weather forecasting in Saudi Arabia by training both the individual neural network and the hybrid network using weather elements that exist in the dataset. The outcome was either rainy or dry, the inputs were appointed to determine correlation coefficient, Root Mean Square Error (RMSE) and scatter index. The paper showed that the hybrid model was better of the individual front grille model (MLP and RBF) and the results were more accurate and had a better learning ability.

In other paper [10], the authors studied a high-precision temperature prediction through complex data for atmospheric. There were two types of weather prediction: dynamic and experimental. They used the Back Propagation Neural Network (BPN) approach and Feed Forward Neural Network and they used randomly weights for all nods to train a data collection using three hidden layers to numerical weather prediction.

According to paper [11], the authors studied forecasting rainfall in India using artificial intelligence techniques to support agriculture and crop multiplication through predicting the precipitation for the next year, they studied a precipitation of historical data and its relationship to the atmosphere using Multiple Linear Regression (MLR) approach. Moreover, the data used is for 30 years from 1973 to 2002 and included data on cloud cover, average temperature and precipitation to Udaipur city, Rajasthan, India. The system was predicting the monthly rainfall quantities, which was very closed to the actual results.

Authors in paper [3] were predicting the weather using arbitrary decision tree algorithm. The elements in the dataset were divided using divide and conquer technique. The data used in this study were obtained from [12] for the city of London. In this study, the weather was predicted by studying satellite data or by the behavior of animals affected by weather changes, weather maps. Finally, using split evaluator, information gain and entropy, they got higher resolution and a small decision tree.

The authors studied in paper [13] a data classification tools and made a detailed comparison between the three tools Decision tree, KNN, and Naive Bayes. They explained the advantages and disadvantages of each one and explained how they are working. Moreover, some examples were given for each type of tools, and applied some examples of weather forecasting. Finally, they concluded that the decision tree was the best and most accurate.

According to the study in [14], authors have tested the applicability of soft computing technique-based rainfall-runoff models (ENN and ANN) to simulate runoff in Bihar. Runoff and antecedent runoff, precipitation, antecedent precipitation over the basin, at three gauging stations in the basin were first identified as appropriate input variables, and then CCF curves at differ time lags were plotted to select the potential input variables. Monthly rainfall data of two stations and discharge data of one station for the period 1986-2014 were utilized as data sets for the development of proposed models. Based on their statistical, it was indices it had been established that ENN outperformed ANN and is more accurate as compared to the traditional ANN method for rainfall-runoff modelling. The results of their study were helpful in selecting the appropriate model for the discharge simulation in Bihar and thereby helping planners for effective flood mitigation.

It can be concluded that there are many researchers have done work related to the weather prediction using machine and deep learning algorithms, but very limited applied their studies for domes specially for mosques. Next section, the methodology of our study will be presented in detail.

III. Methodology

In this paper, Knowledge Database Discovery (KDD) methodology has used to present all the steps required to build up the model from the data collection stage to the preprocessing, cleaning, selecting the data, then choosing the appropriate model and finally evaluating the results. The steps of KDD are mentioned in Figure 1.

A. Selection

In this paper, Data were collected from Kaggle website [12] for the weather in Saudi Arabia, the dataset contains the hourly changing weather from 2017 to 2019 for all the cities in Saudi Arabia. The size of dataset is 249024 records. However, the selected attributes are date, hour, minute, day, temperature, humidity, wind strength, barometer and visibility.

B. Preprocessing

Using the correlation function, the relationship between features has been showed, then some uncorrelated features such as date, time, year, month, day and minute were ignored for the dataset. Al Madina city was chosen as target city. Finally, after cleaning the dataset, 19964 records were left for further processing.

C. Transformtion

In this stage, some other columns such as “state” and “New weather” have been added, and then the state of the weather for thirty six state have converted to (0, 1) and added into “new_ weather” column, and the "state" column has the final state for domes where 0 means the dome is close and 1 means the dome is open. The following Table I. shows the transformation of the weather attributes.

Temperature could change the state of domes, so if the final domes state "new weather" is 1 and the temp degree is more than 16° and lower than 27°, then the value of state column will stay 1 (keep domes open). Otherwise, the value of state will be changed to 0 (close domes). Figure 2 shows the impact of temperature factor on domes state.

D. Modeling

In this section, the weather status and domes state will be predicted using Decision Tree (DT) and k-Nearest Neighbors (k-NN) algorithms. The proposed model can determine the state of the domes through seven factors. When temperature below 16° or more than 27°, the domes must be closed. Otherwise, the domes state must be 1 or 0. As mentioned in Figure 4, case 1 means that domes are open while the case 0 means that domes are close. After determining the state of the domes, the system will give a signal to the Arduino controller device, where the Arduino controller operates the rails to open and close the domes.

Regarding to the air conditioners, if the domes are open, air conditioners will be turned off, but if the domes are open, then the air conditioners will be turned off. Finally, to solve the problem of unexpected rainfall, rain detection sensor will be used, see Figure 3.

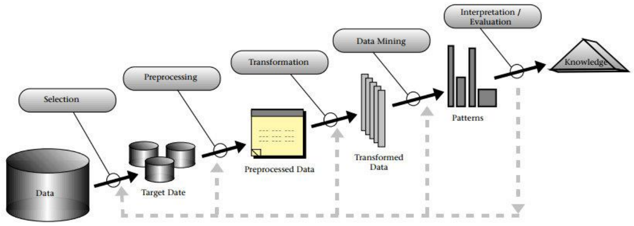


Fig. 1. Knowledge Database Discovery (KDD) Processes [15].

TABLE. I. The Transformation of the Weather Attributes

State	Visibility	Barometer
1	1	Clear
2	0	Sunny
3	1	Passing clouds
4	1	Low level haze
5	1	Scattered clouds
6	1	Partly sunny
7	1	Broken clouds
8	0	Duststorm
9	0	Sandstorm
10	1	Pleasantly warm
11	1	Thunderstorms passing clouds
12	1	Thunderstorms partly sunny
13	1	Thundershowers
14	1	Mostly cloudy
15	1	Thunderstorms Broken clouds
16	1	Thunderstorms Scattered clouds
17	0	Extremely hot
18	1	Mild
19	1	Thunderstorms Partly clouds
20	0	Rain Partly cloudy
21	0	Rain Scattered clouds
22	0	Rain Broken clouds
23	1	Haze
24	1	Overcast
25	1	Dense fog
26	0	Rain passing clouds
27	0	Rain Mostly cloudy
28	0	Rain Partly sunny
29	1	Fog
30	0	Hail Partly sunny
31	1	Thundershowers passing clouds
32	1	More clouds than sun
33	1	Thunderstorms more clouds than sun
34	1	Thunderstorms
35	1	Partly cloudy
36	0	Hail

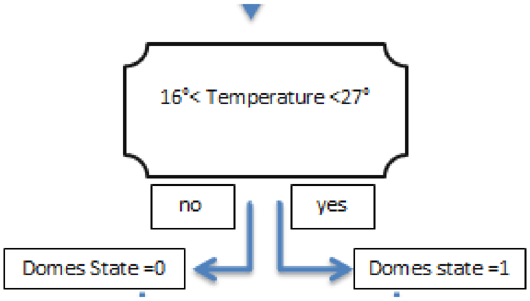


Fig. 2. Impact of Temperature Factor on Domes State.

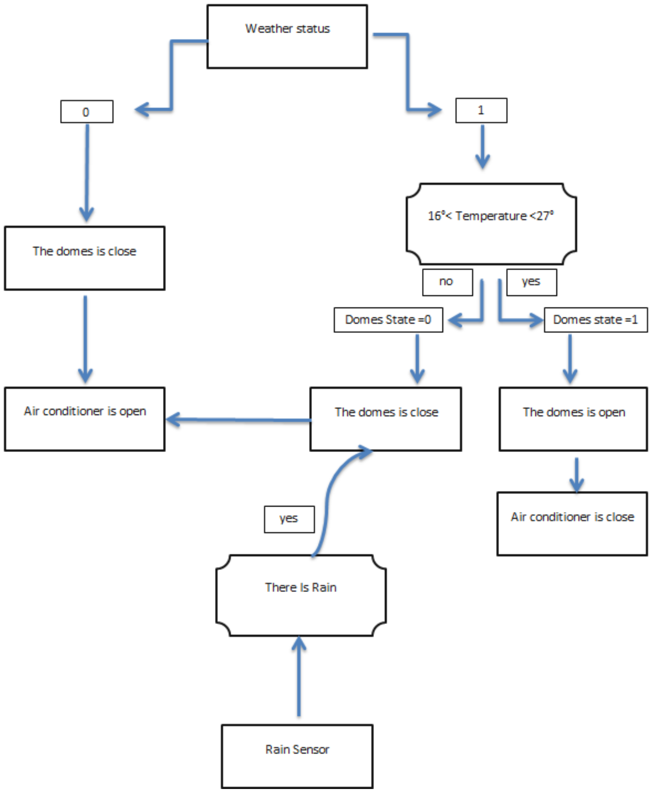


Fig. 3. Proposed Model of Weather and Domes Status.

In this stage, a model will be built and due to the study [13], it has been found that the decision tree and kNN algorithms are most effective than others. Furthermore, both algorithms have high accuracy for weather forecasting, in addition to the speed of training and testing.

Weather forecasting is a complicated process. However, there are two types of weather forecasting techniques, one is the dynamic and the other is experimental. The experimental prediction is used by meteorologists if there are a lot of data and used in a local area, the dynamic prediction is used for broad forecasting and it is not ineffective with short-term.

1) Decision Tree (DT): A decision tree is predictive modeling technique used in classification, clustering and prediction tasks. DT is one of the most common machine learning algorithms and it uses divide and conquer technique to split the problem search space into subsets [16].

Python programming language will be used for the implementation stage. However, the prediction model of the weather and domes status will be built using sklearn with some important libraries and methods such as accuracy_score, train_test_split, DecisionTreeClassifier and panads.

The algorithm aims to divide and distribute the records of dataset into depth-first greedy approach or breadth-first approach, where the structure of the algorithm must consist of root, internal nodes and leaf, where each node refers to a condition on the attribute and uses approach from top to bottom. In addition, the decision tree is quick, simple and easy to understand of representation.

After finishing the preprocess of the previous step, data will be entered to the model to be trained to build the model and then the model will be tested. For this step, the target (output) and the attributes (inputs) have been determined. See Table II.

As mentioned in Table 2, the weather features (“Temp”, “wind”, “humidity”, “hour”, visibility”, “barometer”) will be used as inputs to the model, and “state” will be used as output. Data have divided into two splits, 33% for testing and 77% for training. The max leaf nodes are 50 nodes and random state is 324.

After testing the model, the accuracy has been calculated through the accuracy functions which have resulted 98%. In addition, other evaluation methods have been calculated as well. Table III shows the results of evaluation methods for DT.

2) k-Nearest Neighbors (k-NN): k-NN algorithm is one of the best algorithms for Machine Learning, which is easy to use, and it is introduces an excellent accuracy compared to other algorithms [17]. k-NN is based on calculating the distance between the point required with all points in the neighborhood, and choosing the shortest distance, which depends on the value of k, where k is the number of neighborhoods must be compared with the required point. K-NN is the fastest technology to learning comparing with neural network, decision tree and Bayes networks, but it takes a long time during the classification process and works well on data with multiple classifications.

In this paper, k-NN is used with 141 k’s, where k has calculated by the square root of the data records 19964. Python programming language will be used for the implementation stage. However, the prediction model of the weather and domes status will be built using sklearn with some important libraries and methods such as accuracy_score, train_test_split, KNeighborsClassifier, pandas, matplotlib.

The weather features (“Temp”, “wind”, “humidity”, “hour”, visibility”, “barometer”) have used as inputs to the model, and “state” has used as output. Data has split into 30% for testing and 70% for training, and random state is 101.

After testing the model, the accuracy has been calculated through the accuracy functions which have resulted 95%. In addition, other evaluation methods have been calculated as well. Table IV shows the results of evaluation methods for k-NN.

TABLE. II. SAMPLE OF PREPROCESSED DATA FOR WEATHER FEATURES

State	Visibility	Barometer	Humidity	Wind	Temp	hour	
1	16	1020.0	0.33	0	21	24	0
1	16	1020.0	0.35	9	19	1	1
1	16	1020.0	0.37	11	19	2	2
1	16	1020.0	0.40	7	18	3	3
1	16	1019.0	0.39	0	17	4	4

TABLE. III. THE RESULTS OF EVALUATION METHODS FOR DT

	F1→ 1	F1→ 0	Weighted Avg → F1	MSE	Accuracy
Decision Tree	0.97	0.99	0.98	0.019	0.98

TABLE. IV. THE RESULTS OF EVALUATION METHODS FOR K-NN

	F1→ 1	F1→ 0	Weighted Avg → F1	MSE	Accuracy
k-NN	0.91	0.96	0.95	0.055	0.95

IV. RESULTS INTERPRETATION AND EVALUATION

In this section, the results obtained from the previous sections for both Decision Tree (DT) and k-Nearest Neighbor k-NN algorithms will be interpreted and discussed in this section. In fact, more than one evaluation criteria have been used to evaluate the proposed model. Model evaluation is an integral part of the model development process, which helps to find the best model that represents the data and how well the chosen model will work in the future. The used evaluation measures are described in detail below:

- Accuracy: the most commonly used metric to judge a model. It can be measured according to the percentage of the recognized hand images per the total number of tested hand images. Accuracy can be calculated as following, where TP is the true positive instances, TN is the true negative instances, FP is the false positive instances, and FN is the false negative instances.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

- F1 Score: also called F-measure, considers both the precision and the recall to compute the score. The F1 score is the harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

- Mean Square Error (MSE): measures the average squared difference between the estimated values and true value. It is a risk function, corresponding to the expected value of the squared error loss, always nonnegative, values close to zero are better.

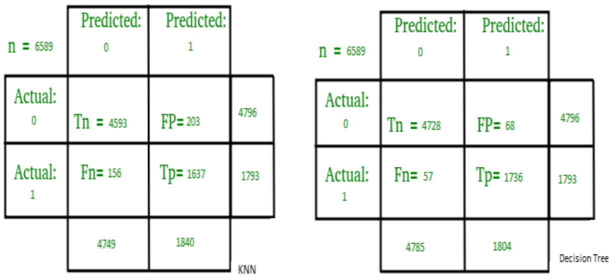
According to the following Table V., it has been found that DT has a higher accuracy with a value 0.98 than k-NN with a value 0.95, and higher F1 function with a value 0.98 comparing with k-NN with a value 0.95.

TABLE. V. COMPREHENSIVE RESULTS INTERPRETATION AND EVALUATION

	F1→ 1	F1→ 0	Weighted Avg → F1	MSE	Accuracy
Decision Tree	0.97	0.99	0.98	0.019	0.98
k-NN	0.91	0.96	0.95	0.055	0.95

In term of Mean Square Error (MSE), DT has lower value 0.019 than k-NN with a value 0.055, which means that Decision Tree method has higher performance comparing with k-NN method to predict the state of the mosque's domes using weather features and outside temperatures.

The following Figure 4 shows the Confusion Matrix for k-NN and DT.



n = 6589	Predicted: 0	Predicted: 1	
Actual: 0	Tn = 4593	FP= 203	4796
Actual: 1	Fn= 156	Tp= 1637	1793
	4749	1840	KNN

n = 6589	Predicted: 0	Predicted: 1	
Actual: 0	Tn = 4728	FP= 68	4796
Actual: 1	Fn= 57	Tp= 1736	1793
	4785	1804	Decision Tree

Fig. 4. Confusion Matrix for k-NN and DT.

After comparing all the results which have been achieved from Confusion Matrix as mentioned in Figure 4 for both models, it has been found that the DT algorithm was better method comparing with k-NN, so it will be used for building the smart domes model.

Given the results that we obtained previously, we can say that we have succeeded in building the smart domes system to the Prophet's Mosque in Saudi Arabia which can be expanded to other mosques all over the world, where we can now control the domes to make decisions automatically using machine learning algorithms considering some significant weather features and outside temperatures. In addition, the problems of ventilation, pollution, spread of germs and inappropriate smells in the Mosques have been solved successfully.

V. CONCLUSION

In this paper, a system of smart domes has been proposed using weather features and outside temperatures. Machine learning algorithms such as k-NN and Decision Tree have been applied on weather features and outside temperature to predict the state of the mosque's domes (open or close). The results of this study are promising and will be helpful for all mosques to use our proposed model for controlling domes automatically.

Due to the difficulty of ventilating the mosques, decreasing the pollution, spread of germs and inappropriate smells in the Mosques, our proposed model will be very good to solve these problems and keep the mosques healthy for worshipers and visitors.

In the future work, some ideas will be applied to solve the problem of people crowding inside the mosque by determining the time for dome to be opened in minutes by specifying the number of worshipers inside the mosque using fuzzy based control, specially that fuzzy systems have been applied in many different industries [18]. In addition to applying some other machine and deep learning algorithms in the future to increase the performance of the model.

References

[1] Deloitte and Dubai Islamic Economy Development Center. https://www.sasapost.com/largest-numbers-of-mosques/.

[2] General Presidency of al Haram Mosque and the Prophet’s Mosque. https://www.gph.gov.sa/ar-sa/MasjidulNabawi/Pages/Building-and-expansion-Nabawi-Mosque.aspx.

[3] Nalanda B Dudde, Dr.S.S. Apte , “Arbitrary Decision Tree for Weather Prediction” , International Journal of Science and Research (IJSR) , Vol. 5 , Issue. 3,pp. 87-89, doi:10.21275/v5i3.nov161774. 2016.

[4] General Authority for Statistics ‘kingdom of Saudi Arabia’. https://www.stats.gov.sa/ar/news/203.

[5] M.Kannan, S.Prabhakaran, P.Ramachandran, ”Rainfall Forecasting Using Data Mining Technique”, International Journal of Engineering and Technology Vol. 2, no. 6,pp. 397-401, 2010.

[6] A. Blasi, “Performance Increment of High School Students using ANN Model and SA Algorithm”. Journal of Theoretical & Applied Information Technology, 95(11):2417-2425. 2017.

[7] B. M. Gupta and S. M. Dhawan, “Deep Learning Research: Scientometric Assessment of Global Publications Output during 2004 -17,” Emerging Science Journal, vol. 3, no. 1, p. 23, Feb. 2019. doi:10.28991/esj-2019-01165.

[8] S. Ghorbani, M. Barari, and M. Hosseini, “A Modern Method to Improve of Detecting and Categorizing Mechanism for Micro Seismic Events Data Using Boost Learning System,” Aug. 2017. doi:10.20944/preprints201708.0072.v1.

[9] Saba, T., Rehman, A., & AlGhamdi, J. S. “Weather forecasting based on hybrid neural model”. Applied Water Science, 7(7), 3869–3874, 2017. doi:10.1007/s13201-017-0538-0.

[10] Ch.Jyosthna Devi, B.Syam Prasad Reddy, K.Vagdhan Kumar,B.Musala Reddy,N.Raja Nayak , “ANN Approach for Weather Prediction using Back Propagation ” , International Journal of Engineering Trends and Technology , Vol. 3 , Issue. 1,pp. 19-23, 2012.

[11] Nikhil Sethi, Dr.Kanwal Garg , “Exploiting Data Mining Technique for Rainfall Prediction”, International Journal of Computer Science and Information Technologies, Vol. 5, No. 3,pp. 3982-3984, 2014.

[12] Saudi Arabia weather history data. Kaggle Website: https://www. kaggle.com/esraamadi/saudi-arabia-weather-history.

[13] Sayali D. Jadhav, H. P. Channe , “Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques ”, International Journal of Science and Research (IJSR), Vol. 5 Issue. 1, pp. 1842-1845, 2016.

[14] S. Kumar, T. Roshni, and D. Himayoun, “A Comparison of Emotional Neural Network (ENN) and Artificial Neural Network (ANN) Approach for Rainfall-Runoff Modelling,” Civil Engineering Journal, vol. 5, no. 10, pp. 2120–2130, Oct. 2019. doi:10.28991/cej-2019-03091398.

[15] J. Han, M. Kamber, and J. Pei, “Classification,” Data Mining, pp. 327–391, 2012. doi:10.1016/b978-0-12-381479-1.00008-3.

[16] F. LI, Y. LI, and C. WANG, “Uncertain data decision tree classification algorithm,” Journal of Computer Applications, vol. 29, no. 11, pp. 3092–3095, Dec. 2009.

[17] O. Kramer, “K-Nearest Neighbors,” Intelligent Systems Reference Library, pp. 13–23, 2013.

[18] A. Blasi, “Scheduling food industry system using fuzzy logic”. Journal of Theoretical & Applied Information Technology, 96(19): 6463-6473. 2018.

Accident Detection and Disaster Response Framework Utilizing IoT

Shoaib ul Hassan[1], Jingxia CHEN[2], Ali Akbar Shah[4]
Department of Computer Science and Technology
Shaanxi University of Science and Technology
Xi’an, China

Tariq Mahmood[3]
Department of Computer Science and IT
University of Sargodha
Bhakkar, Pakistan

***Abstract*—The internet of things (IoT) leads the noteworthy edges above customary information and communication technologies (ICT) for Intelligent Transportation Systems (ITS). The progression in the transportation system, the increment in vehicles and the accidents happened on the roads are cumulative up to an alarming situation. Additionally, 1.256 million people expire by the road bumps every year and it is very problematic to find the precise accident location of the user. If an accident occurs, the survival rate of the victim increases, if he is given instantaneous remedial assistance. You can provide remedial assistance to the victim only when you identify the precise location of accident. The main persistence of this system is to identify an accident and find the location of the user. After tracing the location, the system will search nearby hospitals for remedial treatment. System will send a message that contains user's current location, to the nearby hospitals in case of an emergency. System will acquire recommended contacts from the cloud and also send message to them for user’s support by using API. If the user is safe, he can cancel the message that is being sent to the nearest hospital and the recommended contacts. This system will help the users in saving their lives within minimal time.**

Keywords*—Internet of things (IOT); accident detection; nearby places; nearby hospitals; cloud computing; intelligent transportation systems; information and communication technologies

I. Introduction

Nowadays, researchers have described a lot of benefits of using smart cell phones in accidents and report systems in numerous articles. The main reason of using such type of software in a smart cell phone is because of their installation in vehicle is much more expensive than a smart cell phone. Smart cell phones are always present with their possessors and can signify an accident, even though if the vehicles somehow are not engaged in the accident. Plus smart phone uses integrated sensors [17] (such as accelerometer, GPS, gyroscope, etc.) which help in getting a rich variety of data.

Vehicle accidents are an unavoidable issue in your routine life. The recklessness of driver causes numerous vehicle accidents. Additionally, these accidents produce a financial and social damage as well as in worst case scenario, maybe the end of a precious human life [1]. The security of driver and traveler can be endangered due to different factors that lead up to a mishap. Besides this, there are huge time difference between the hour of mishap and when the crisis administrations arrive at the point of accident. Accidents are identified by utilizing three sensors for example accelerometer, force resistive sensor and gyroscope to get precise outcomes [2]. More often, we can't find the accurate location of the accident when we don't have the exact idea where the accident has occurred. We can utilize the essential microcontroller AT89S52 and assembly programming for improved precision and GPS and GSM module to follow the vehicle anyplace on the globe [3]. Moreover, report it to the closest remedial care center [4] to save precious human life.

We can lessen the human demise proportion and give appropriate assistance in the most barren regions by presenting a solution to detect the location of accidents [5]. The data is transmitted to the recommended contacts immediately when the accident happens.

The Internet of Things (IoT) has seen development in the past few years, with upgrades in a few unique applications in the armed, naval, intelligence transportation and many other fields which belongs to the safety and convenience of human beings. In spite of these facts the IoT brings huge points of interest over customary information and communication technologies (ICT) for Intelligent Transportation Systems (ITS), anyhow accidents are quiet very common and are increasing day by day [19]. These accidents can be recognized by utilizing different applications or some other vehicle communication system. Although, the algorithms generated by the machine can straightforwardly find out the accidents [20] by using the speed of the vehicle or by some sudden jerks. Accidents or different occurrences can be viewed as irregularities in rush hour jams, information and machine algorithms [21, 6] can be utilized to distinguish these exceptions. Upon exposure to the unwanted event, the incoming drivers can be warned about that mishappening, and this will support the users take certain measures to avoid greater disaster.

GPS works as a vibrant piece of the automobile that gives how far the vehicle is moving, at what time and how close it is to its destination. This system helps to identify an accident from the location of the automobile by using GPS speed information and corresponding maps algorithms and send out the accident site to an Alert Service Center [3, 7]. The position information will be utilized in the corresponding map algorithms to find out the automobile position. Whenever, the speed will slow down beneath the predetermined safe-limit, the system will identify the case as an accident. Also, certain accidents can be distinguished utilizing the GPS and GSM technologies, vehicular ad-hoc networks [8] (VANET) and versatile framework.

A couple of prevailing works have exhibited an accident finding system on an Android smart cell phones. When the accident is recognized, a warning is sent to every vehicle in the vicinity, in addition to that, a message or voice call is also transmitted to the user’s recommended contacts. Moreover, to regulate the false-positives, these warnings are only conveyed if the user is unable to interfere with the commencement instruction, which starts with the accident detection algorithm [9, 10]. This will quickly alert the closest police headquarters as well as the remedial clinic and send emergency messages for assistance.

In this article, we struggled in recognizing the accident and finding out the precise location of the user. Relevant to the subscribed location, the system will scan close by emergency clinics and hospitals for remedial treatment. System will deliver message that holds client current location to the nearby emergency clinic for assistance and remedial treatment. System will acquire the contacts of recommended personnel from the cloud and will send the message to client's supporter for assistance that holds user present location. On the other hand, if the drive is safe, he can also cancel the message that is being sent to the closest remedial clinic and recommended contacts. This system will assist the user in saving their life. Clients could be efficiently found and get remedial treatment in minimal time.

This article is prearranged as follows. Segment 2 describes material and methods portion. Segment 3 describes the experiments and the results. Segment 4 examines the results. Segment 5 offers some concluding observations and suggestions for future work.

II. Material and Methods

A. Android Studio

In this article, we utilize Android Studio [11-12] to develop this system and perform experiments. Android Studio is the state Integrated Development Environment (IDE) used for improvement in an Android application. It is pattern on the IntelliJ IDEA, the Java coordinated advancement requirement for programming, and have all the code modifying and engineer devices. To help out the application progression through the Android operational framework, Android Studio utilizing a Gradle-based framework, emulator, code formats, and GitHub joining. It is much more easy, flexible and using Android Studio we can even develop complex android based applications. It also provides us built-in debugging and testing features to make our code clean and to run our applications smoothly. The individual undertaking in Android Studio has minimum one system with source code and asset report. These techniques incorporate the Android application modules, Library modules, and the Google App Engine modules. The Android Studio is using an Instant Run feature to push code and resource changes to currently running application on real device or android emulator. A code proofreader encourages the programmer with composing code and delicate code execution, scattering, and broke down. Applications developing in the Android Studio are proceed in the Google Play Store for further arrangement into the APK group for consistence. Android Studio provide us to generate Android App Bundle file to publish our application on Google Play Store. Android App Bundle is new format that includes all code, resources, files, build settings etc. of your application. When a user tries to download applications form Google Play Store, it generates differ APK and deliver to the user according to his device.

Android Studio 3.5 as shown in Fig. 1 has auto-recommended memory settings. Truth be told, when the Android Studio perceives that your task requires more RAM, it is naturally increment in the memory store size, however it likewise tells the android application developer for the equivalent.

Fig. 1. Android Studio 3.5.

B. Java and XML

Java is a far-reaching programming language that is plainly anticipated for the use of communicating the requirement of web. It is a supremely stylish programming language for Android smart cell phone applications. Furthermore, it is the best upheld in the advancement of gadgets and the web of things. Java remained adjusted to have the aspect and impression of the C++ language, however it is simple to develop and execute the items arranged programming model. Java can be developed to create entire applications that might keep running on an individual PC or be circulated among servers and clients in the system. Similarly, it can be used to assemble the minute application module or gadget for its usage as a major aspect of the site page. Java is platform of independent language which means that Java can be installed and run on almost every operating system. More than 3 billion devices are using Java these days and it is growing day by day. Java is pure object-oriented language and it also supports multithreading. Finally, Java is much more secure than other different languages.

Extensible Mark-up Language (XML) is the language that is much like HTML, but XML is much more flexible than HTML. XML provide us to make user defined tags. XML is designed to store and ship information and it is arranged to be self-descriptive. That characterizes a lot of norms for encoding, archives is an example that is both comprehensible and machine-discernible. The purpose for creating XML emphasizes on its effortlessness, sweeping report, and its easiness for the usage on the Internet. It is a textual data format with Unicode providing assistance from various human languages. XML is intensively utilized as an arrangement for report stockpiling and handling, both on the web and disconnected PC depending on global measures. Forward and in reverse resemblance is generally simple to keep up in spite of changes in DTD (Document Type Definition) or Schema. It very well may be refreshed steadily. Although, plan of the XML accentuation on records, the language is generally utilized for the outline of discretionary information structures, e.g. utilized in web administrations.

C. Device Shake Detection

Motion sensor devices are helpful in inspecting device movement, e.g. incline, wobble, pivot or swipe. The motion is generally a perception through user contribution (for instance, user guiding a vehicle in the game or a client controlled ball in a game), however it can likely be a sensation of physical condition where the device is stationary (for instance, it will move with you while you are travelling on your vehicle) [13]. In the principal circumstance, you are checking measures, in esteem to the device's indication or your system's alert. In the 2nd case you're observing mobility with respect to the world's casing of reference. The mobile sensors without someone else's input are not normally utilized as screen gadget position, however it can be functional with different sensors, e.g. the Geo-magnetic field sensor, to determine the device's location or shake detection. In order to avoid unnecessary false alarm, this framework check the intensity of vibration and threshold value of force. If the threshold value of force is less than 10000 Newton (N), our framework does not detect any accident. If the threshold value of forces is greater than 10000 Newton and user is safe, then he can also cancel the alarm and help message. Fig. 2 demonstrates the real-world implementation that device shake sensors then it works when accident is detected.

D. Firebase Realtime Database

In this paper, we used the firebase real-time database as shown in Fig. 3 for the storage and recovering data [14]. Google Firebase is a Google-supported application advancement programming that enables designers to create iOS, Android and Web applications. Firebase gives apparatus to following investigation; revealing and fixing application crashes, making advertising and analyzing items. The Firebase Realtime Database is a cloud-facilitated NoSQL database that enables information to be put away and adjusted between clients progressively. The information of all customers is synchronized continuously; this is advantageous for the case when the application is disconnected. Firebase Cloud Messaging (FCM) is a cross-stage informing instrument that lets venture securely get and convey messages on iOS, Android and the web at no expense. Firebase Authentication makes it simple for engineers to construct secure confirmation frameworks and amplifies the sign-in and mix involvement for clients. This element offers a total personality arrangement, supporting email and secret word accounts, telephone authentication, just as Google, Facebook, GitHub, and Twitter login.

Here, we used Firebase real-time database [14] to store the supporter contact details and accident location, where we verified 32 casual accident locations and the database model was designated as demonstrated in Fig. 4.

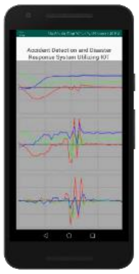

Accident Detection and Disaster Response System Utilizing IOT

Fig. 2. Device Shake Sensor.

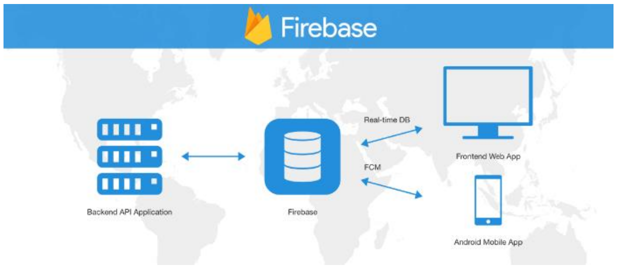
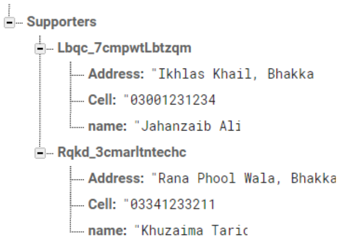


Fig. 3. Firebase Real-Time Database.



```
Supporters
    Lbqc_7cmpwtLbtzqm
        Address: "Ikhlas Khail, Bhakka
        Cell: "03001231234
        name: "Jahanzaib Ali
    Rqkd_3cmarltntechc
        Address: "Rana Phool Wala, Bhakka
        Cell: "03341233211
        name: "Khuzaima Taric
```

Fig. 4. Firebase Real-Time Database for Support Contact Information.

E. Google Places API

Google Places API is the facility that is presented by Google which processes data about your closest places by examining the latitude, longitude and range of territory. It is a service that provides information about location [15]. You can include or remove a spot from their "places service" as well. Google Places API provides many built-in features to facilitate its users. The arrival of Android places API encourages the information access as well as maintains a strategic distance from the engineers to monitor latitude and longitudes [22]. Prior to access place information in Android, one needs to get all the data from a web administration by passing different parameters like latitude and longitudes to it. The API utilized around past was Google Maps API, but the new Google places API is amazing enough to recognize your current location, nearby places data and recover all the travel information arithmetically. We use Google Places API in our framework to get exact location of the user when he met with accident. Fig. 5 demonstrate the key code segment, how we are getting data of nearby hospitals from the firebase database.

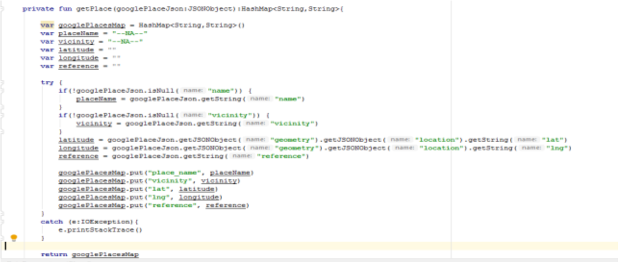
F. GSM and GPS

GSM is developed to send and gather information from a focal unit through an information call. GPS is a satellite way system that attires zone and time data in all climate circumstances to the client and used to decide the ground area of an article. There are various electronic gadgets dependent on GPS, while GSM innovation is typically utilized by cell phone gadgets. On the off chance that your cell phone has a GPS chip, it must be on the off chance that you need to find it utilizing real GPS satellites [23]. On the off chance that your cell phone doesn't have a GPS chip, or if it is killed, at that point GSM restriction will triangulate the position utilizing the three nearest GSM base stations. Limitation of utilizing GSM [14] will work to convey your general position, and it likewise consume your phone battery, however it can’t give exact position data.

To find out the straight path in a graph we utilized the Dijkstra’s algorithm, as shown in Fig. 6. GPS is utilized in Dijkstra’s algorithm to acquire the existing position of respective node. Distance may also be calculated from its position. The essential part of the algorithm is to use what controls to traverse the functionality at Google Maps, Apple Maps, at this juncture, Open Street Map and any additional ordinal map that perhaps you may use. Yes, it is not quite the same algorithm that controls the navigation application nowadays, but exploration and additional algorithms are an addition of the unusual Dijkstra’s algorithm. Dijkstra’s Algorithm is used by Google maps to find the shortest path and nearby locations from specific place. It’s also used to calculate the distance between edges. By using Dijkstra’s Algorithm, Google places API suggests us nearby places/hospitals from the accident location for medical treatment. Fig. 6 demonstrates the Dijkstra’s algorithm is utilized to locate the direct track in the graph.

G. Accident Data Interpretation

The data which is sent as Short Message Service (SMS), will be acknowledged by nearby hospital and user’s supporter [7]. A suitable system will be inscribed so that the accident site is inevitably acquired and sent to the nearby hospital and user’s support. If the driver is safe or accident is falsely detected, he can cancel the SMS. Fig. 7 demonstrate the flow chart of the system.



```
private fun getPlace(googlePlaceJson:JSONObject):HashMap<String,String>{

    var googlePlacesMap = HashMap<String,String>()
    var placeName = "--NA--"
    var vicinity = "--NA--"
    var latitude = ""
    var longitude = ""
    var reference = ""

    try {
        if(!googlePlaceJson.isNull( name: "name")) {
            placeName = googlePlaceJson.getString( name: "name")
        }
        if(!googlePlaceJson.isNull( name: "vicinity")) {
            vicinity = googlePlaceJson.getString( name: "vicinity")
        }
        latitude = googlePlaceJson.getJSONObject( name: "geometry").getJSONObject( name: "location").getString( name: "lat")
        longitude = googlePlaceJson.getJSONObject( name: "geometry").getJSONObject( name: "location").getString( name: "lng")
        reference = googlePlaceJson.getString( name: "reference")

        googlePlacesMap.put("place_name", placeName)
        googlePlacesMap.put("vicinity", vicinity)
        googlePlacesMap.put("lat", latitude)
        googlePlacesMap.put("lng", longitude)
        googlePlacesMap.put("reference", reference)
    }
    catch (e:IOException){
        e.printStackTrace()
    }

    return googlePlacesMap
```

Fig. 5. Getting nearby Hospitals Data from Accident Location.

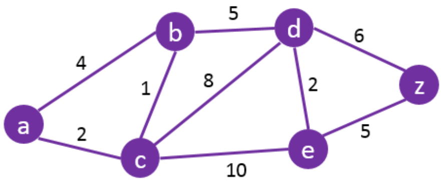


Fig. 6. Dijkstra's Algorithm is Utilized to Locate the Direct Track in a Graph.

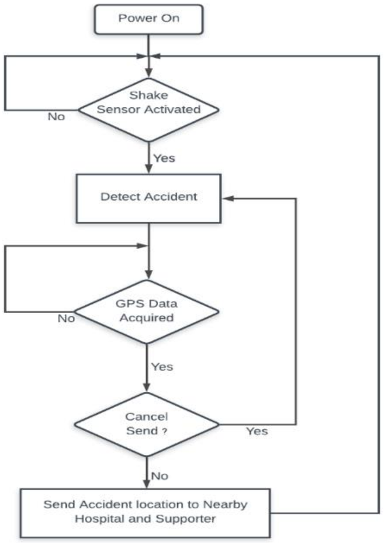


Fig. 7. Flowchart of the System.

III. Experiments and Results

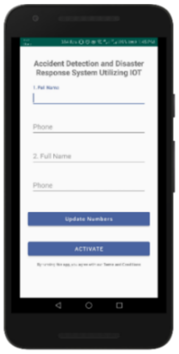
Here, in this fragment, the execution of the anticipated system and precision of this recognition is to be examined. To achieve the assessment procedure [16], we have smart cell phone which is always in the pocket of the user. It is renowned that the number of individuals engaged in this exercise and those who are involve in the test panel both are four. However, the four individuals in the test group and that in the exercise panel both are different. The data assimilated from the four individuals in the exercise panel is used to find out the threshold settings, while the others, four persons who are in the test panel are for efficiency validation. Furthermore, the projected accident detection and disaster response framework is applied in the Huawei Mate 10 Lite smart cell phone with the following characteristics.

- CPU: HI Silicon Kirin 659 Octa-core (4x2.36GHz)
- RAM: 4GB
- ROM: 64GB
- OS version: Android 8.0

A. Real-World Implementation

On this stage, application will get supporters information and save all the information on the cloud for future use. At this stage, user must provide 2 emergency recommended contact numbers [24, 25]. If user already saved the emergency contact number on the cloud, there is no need to save number on the cloud every time they access the system.

He can update the emergency contact numbers at any time by giving new numbers by clicking on “update numbers” button. Fig. 8 demonstrates the real-world implementation that a user can save his information and emergency contact numbers on the cloud.



Accident Detection and Disaster Response System Utilizing IOT

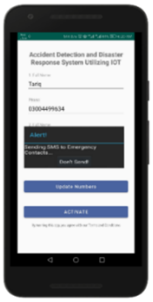
Fig. 8. Real-World Implementation.

After updating all the information, user should activate the application by clicking on *"*activate*"* button. This action is required for the first time only. After activation, system will automatically detect accidents and show an alert box to the user before sending alert message to the recommended contacts and nearby hospitals. Fig. 9 demonstrates the real-world implementation that the message will be sent to the nearby hospitals for emergency help.

If the driver is safe, he can cancel the alert box which stops the deliverance of the *"*help message*"*, otherwise system will send *"*help message*"* to nearby hospital and recommended contacts. Fig. 10 demonstrates the message sending process to nearby hospitals after getting nearby hospitals data.

B. Information Transfer and Reaction Time

Fig. 11 demonstrates the information transfer and reaction time on the cloud. Time taken by four different android devices to transfer data to the cloud, in milliseconds (MS), to their reaction time on the cloud, with respect to the incident and recommended contact's data.



Accident Detection and Disaster Response System Utilizing IOT

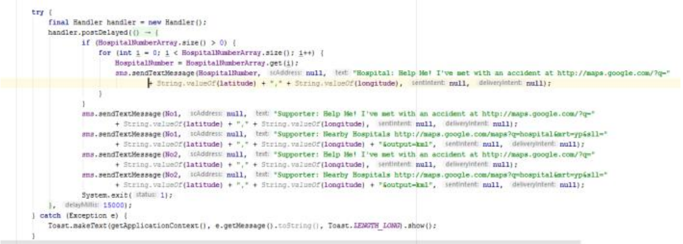
Tariq

03004499634

Alert!

Sending SMS to Emergency Contacts...

Fig. 9. Accident is Detected, and SMS is being sending to nearby Hospital and Supporter.



```
try {
    final Handler handler = new Handler();
    handler.postDelayed(() → {
            if (HospitalNumberArray.size() > 0) {
                for (int i = 0; i < HospitalNumberArray.size(); i++) {
                    HospitalNumber = HospitalNumberArray.get(i);
                    sms.sendTextMessage(HospitalNumber,  scAddress: null,  text: "Hospital: Help Me! I've met with an accident at http://maps.google.com/?q="
                            + String.valueOf(latitude) + "," + String.valueOf(longitude),  sentIntent: null,  deliveryIntent: null);
                }
            }
            sms.sendTextMessage(No1,  scAddress: null,  text: "Supporter: Help Me! I've met with an accident at http://maps.google.com/?q="
                    + String.valueOf(latitude) + "," + String.valueOf(longitude),  sentIntent: null,  deliveryIntent: null);
            sms.sendTextMessage(No1,  scAddress: null,  text: "Supporter: Nearby Hospitals http://maps.google.com/maps?q=hospital&mrt=yp&sll="
                    + String.valueOf(latitude) + "," + String.valueOf(longitude) + "&output=kml",  sentIntent: null,  deliveryIntent: null);
            sms.sendTextMessage(No2,  scAddress: null,  text: "Supporter: Help Me! I've met with an accident at http://maps.google.com/?q="
                    + String.valueOf(latitude) + "," + String.valueOf(longitude),  sentIntent: null,  deliveryIntent: null);
            sms.sendTextMessage(No2,  scAddress: null,  text: "Supporter: Nearby Hospitals http://maps.google.com/maps?q=hospital&mrt=yp&sll="
                    + String.valueOf(latitude) + "," + String.valueOf(longitude) + "&output=kml",  sentIntent: null,  deliveryIntent: null);
            System.exit( status: 1);
    },  delayMillis: 15000);
} catch (Exception e) {
    Toast.makeText(getApplicationContext(), e.getMessage().toString(), Toast.LENGTH_LONG).show();
}
```

Fig. 10. Key Code Segment to send Message to nearby Hospital and Supporter after Detection of Accident.

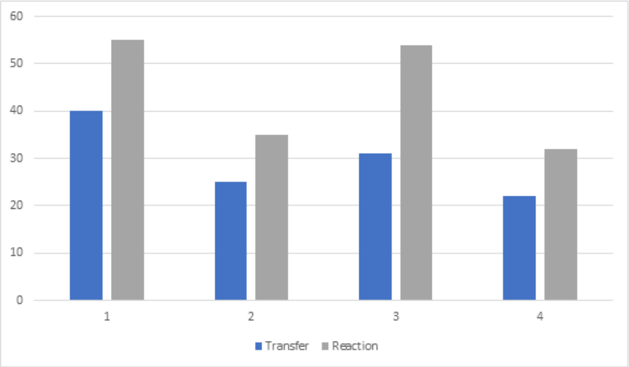


Fig. 11. Information Transfer and Reaction Time.

C. Average Information Transfer and Reaction Time

Average information transfer and reaction done in milliseconds (MS) from online storage. Fig. 12 demonstrates the information transfer and reaction time from online storage.

D. Comparison

Our purposed system provides numerous features. The difference of our projected system is shown in Table 1. Our proposed system is compared with other studies Paper A [17] and Paper B [18].

The purposed framework gives numerous highlights. The correlation of our proposed framework is shown in Table 1. The proposed framework is compared with different papers (Paper A [17] and Paper B [18]).

A1 = Accident alert

A2 = Supporter notification

A3 = Data storage on cloud

A4 = Scalability

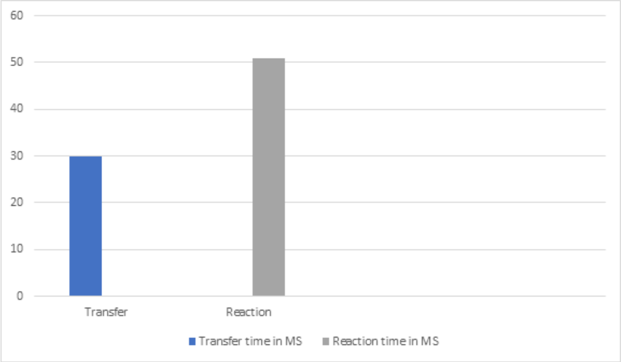


Fig. 12. Average Information Transfer and Reaction Time.

TABLE. I. COMPARISON AND PERFORMANCE OF ACCIDENT DETECTION AND DISASTER RESPONSE FRAMEWORK UTILIZING IOT WITH OTHER SYSTEMS

Feature	Paper A	Paper B	Our System
A1	YES	YES	YES
A2	NO	NO	YES
A3	NO	NO	YES
A4	NO	YES	YES

IV. Discussion

According to the victim's opinion, in case of a mortal accident, the injured person commonly is not capable to contact the rescue by himself, especially in those areas which are not located in the city or not much populated. In these circumstances, the intended system will robotically detect the accident and will send a message that contains user′s current location to nearby hospital for remedial treatment. Sending a disaster alert is quite easy and suitable because all required functionalities squat together. In case of other emergencies, the system also provides capabilities to send request to the anticipated emergency service. Whereas, according to the responder's opinion towards causality, the system will display the precise site of the causality, and it is very beneficial to reach at the location in a minimal time and reduce the response time as well. So that, the authorities will be capable to trail the victims in real time and relief them as rapidly, ensuring in an effective usage of the assets of emergency services.

For instant, let us discuss the A1 performance of the system in previous papers. In paper A and B there are no accident force measurement, but in this article if the force is less than 10000 N, it would not be considered as an accident. In paper B, it takes 2 to 3 minutes to send a voice or text message to the hospital and desired contact but this system send message alert in 15 seconds. Our application is more user friendly than other applications, as it does not require log-in and easy to use. Moreover, other papers don't have support notification features & have not any database attached with their systems.

This study is based upon the shake detection sensors and shows that accidents can be detected by using shake sensors. The person who is using a vehicle can use projected system throughout their journey, and if there are some accident, the system will get the precise accident location by GPS and search for nearby hospitals by *"*Nearby Places API*"* and send all the information to the nearly hospitals as well as to the user's recommended contacts. Basically, when the user falls down by some means, the shake sensor will get some information from cell phone. This is due to the reason that when the cell phone having this system is dropped, the system's shake sensor feels it seriously and becomes activated due to this jerk. Since noticing the normal state as an accident is more desirable than identifying the accident as a normal state, so this can be clinched that the system represents an improved work, and this can be utilized in case of an accident exposure. When there is no accident, the system will detect it as an accident more frequently than vice versa.

V. Conclusion

The Internet of things (IoT) is a system of interrelated computing devices, mechanical and digital machines provided with unique identifiers (UIDs) and the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction. IoT leads the noteworthy edges above customary information and communication technologies (ICT) for Intelligent Transportation Systems (ITS). The progression in the transportation system, the increment in vehicles and the accidents happened on the roads are reaching to an alarming situation. In this paper, we configure a system to identify a road accident and to find the particular site of the accident precisely. To find the precise site of the user, the application uses Global Positioning System. After getting the site of the user, the system will locate the nearby hospitals with the support of Google nearby Places (API) for remedial treatment. System will directly transmit messages containing the user's present site to a nearby hospital for remedial help. The system will get cell numbers of the recommended contacts from the cloud and send a *"*help message*"* containing the user's current location to the contacts for assistance using APIs. If the driver is unharmed, he or she may cancel the message that is being sent to the nearby hospital and recommended contacts. This system will help in saving the life of the users and make secure them from any accidental risk. The user can get remedial care in a minimal time and the user can easily be treated without any further due.

VI. Authors' Contributions

Jingxia CHEN and Tariq Mahmood carried out the best way in making this system for the security purpose as well as to secure the lives of the human beings. I'm personally very thankful to J.C. and T.M. who delivered their best to fulfill the requirements of my idea in a very efficient way and in syndicating the data. A.A participated in experiment design and manuscript drafting. All authors read and are agreed with the final document.

Acknowledgments

We would like to thank the members in this study and the research staff in the Department of Computer Science and Information Technology, University of Sargodha, Sub Campus Bhakkar in Pakistan for their assistance. We would also like to acknowledge the help given by Professor *Jingxia CHEN*, Department of Electrical and Information Engineering, Shaanxi University of Science and Technology, China for her guidance in the research.

References

[1] D. Dalvi, V. Agarval, S. Bansod, A. Jadhave, Prof. M. Shahakar, on Android Application for Automatic Accident Detection, IJARIIE 3 (2017) 735–739.

[2] P. Thakur, S. Singh, G. Shukla, T. Bhutani, S. Negi, on Automatic Accident Detection and Notification System, International Journal of Latest Technology in Engineering, Management & Applied Science 8 (2018) 167–170.

[3] N. H. Sane, D. S. Patil, S. D. Thakare, A. V. Rokade, on Real Time Vehicle Accident Detection and Tracking Using GPS and GSM, International Journal on Recent and Innovation Trends in Computing and Communication 4 (2016) 479–482.

[4] P. Javale, S. Gadgil, C. Bhargave, Y. Kharwandikar, on Accident Detection and Surveillance System using Wireless Technologies, IOSR Journal of Computer Engineering 6 (2014) 38-43.

[5] P. Kaladevi, T. Kokila, S. Narmatha, V. Janani, on Accident Detection Using Android Smart Phone, International Journal of Innovative Research in Computer and Communication Engineering 2 (2014) 2367–2372.

[6] N. Dogru, A. Subasi, on Traffic Accident Detection Using Random Forest Classifier, IEEE (2018) 40–45.

[7] M. S. Amin, M. A. S. Bhuiyan, M. B. I. Reaz, S. S. Nasir, on GPS and Map Matching Based Vehicle Accident Detection System, IEEE 12 (2013) 520–523.

[8] U. Khalil, T. Javed, A. Nasir, on Automatic Road Accident Techniques: A Brief Survey, IEEE 17 (2017).

[9] B. Fernandes, V. Gomes, J. Ferreira, A. Oliveira, on Mobile Application for Automatic Accident Detection and Multimodal Alert, IEEE (2015).

[10] A. B. Faiz, A. Imteaj, M. Chowdhury, on Smart Vehicle Accident Detection and Alarming System Using a Smartphone, 1st International Conference on Computer & Information Engineering, IEEE (2015) 66-69.

[11] A. Singh, S. Sharma, S. Singh, on Android Application Development using Android Studio and PHP Framework, International Journal of Computer Applications (2016) 5-8.

[12] N. Verma, S. Kansal, H. Malvi, on Development of Native Application Using Android Studio for Cabs and Some Glimpse of Cross Platform Apps, International Journal of Applied Engineering Research (2018) 12527-12530.

[13] Y. J. Lee, on Detection of Movement and Shake Information Using Android Sensor, Advance Science and Technology Letters (2015) 52-56.

[14] A. Khan, F. Bibi, M. Dilshad, S. Ahmed, Z. Ullah, H. Ali, on Accident Detection and Smart Rescue System using Android Smartphone with Real-Time Location Tracking, International Journal of Advance Computer Science and Applications (2018) 341-355.

[15] Places API – Overview | Places API | Google Developers, https://developers.google.com/places/web-service/intro, Accessed Sep 19, 2019.

[16] L, J. Kau, C. S. Chen, on A Smart-Phone Based Pocket Fall Accident Detection, Positioning and Rescue System, IEEE Journal of Biomedical and Health Informatics (2014).

[17] A. Fanca, A. Puscasiu, H. Valean, S. Folea, on A Survey on Smartphone-Based Accident Reporting and Guidance System, International Journal of Advance Computer Science and Applications (2018) 409-414.

[18] A. Fanca, A. Puscasiu, H. Valean, S. Folea, on Accident Alert Using IOT and Android Application, International Journal for Research in Applied Sciences & Engineering Technology (2018) 1315-1319.

[19] D. Genoud, V. Cuendet, J. Torrent, on Soft Fall Detection Using Machine Learning in Wearable Devices, IEEE 30th International Conference on Advance Information Networking and Applications (2016) 501-505.

[20] Ekaterina Gurina, on Application on Machine Learning to Accident Detection at Directional Drilling, Journal of Petroleum Science and Engineering (2019).

[21] C. Liao, G. Shou,Y. Liu, Y. Hu, Z. Guo, on Intelligent Traffic Accident Detection System Based on Mobile Edge Computing, IEEE International Conference on Computer and Communications (2017) 2110-2115.

[22] M. Diaz H K. Barberan C, D. Marttinez-M, G. Lopez F, on Offline Mobile Application for Places Identification with Augmented Reality, IEEE (2017) 261-264.

[23] F. Fang, Z. Ding, on High Precision and Accident Detection System for Vehicles in Traffic Tunnel, IEEE 2nd International Conference on Electronic Information and Communication Technology (2019) 419-425.

[24] F. Aloul, I. Zualkernan, R. Abu-Salma, H. Al-Ali, M Al-Merri, on iBump: Smartphone Application to Detect Car Accidents, Computers and Electrical Engineering 43 (2015) 66-75.

[25] S. Madansingh, T. A. Thrasher, C. S. Layne, B. C. Lee, on Smartphone Based Fall Detection System, 15th International Conference on Control, Automation and System (ICCAS 2015) 370-374.

APPENDIX: ABBREVIATIONS

IoT: Internet of Things; CC: Cloud Computing; GPS: Global Positioning System; GSM: Global System for Mobile; ITS: Intelligent Transportation System; ICT: Information and Communication Technologies; VANET: Vehicular Ad-hoc Networks; SMS: Short Message Service; N: Newton. API: Application Programming Interface.

AUTHORS' PROFILE

Shoaib ul Hassan got his B.S. degree from Department of Computer Science and Information Technology in University of Sargodha, Sub Campus Bhakkar in Pakistan in 2018 and doing his MS degree from Department of Computer Science and Technology in Shaanxi University of Science and Technology in China. His research interest is focused on Internet of Things, Android Operating System, Blockchain and Cloud Computing.

Jingxia Chen got her B.S and M.S degrees from Department of electrical and information engineering in Shaanxi University of Science and Technology in China in 2002 and 2005, respectively. During 2013-now, studied for Ph.D. degree in School of Computer Science and Engineering, Northwestern Polytechnical University in China. Now she also works as an associate professor in Department of Electrical and Information Engineering, Shaanxi University of Science and Technology in China. Her research interest is focus on Machine learning and pattern recognition, EEG signal processing and event detection, and deep learning.

Tariq Mahmood got his B.S degree from Department of Computer Science and Technology in University of Sargodha, Sub Campus Bhakkar in Pakistan in 2018 and doing his MS degree from Department of Computer Science in Qurtuba University of Science and Technology in Pakistan. He works as a Visiting Lecturer in Department of Computer Science and Information Technology, University of Sargodha, Sub Campus Bhakkar in Pakistan. His research interest is focused on Internet of Things, Android Operating System and Cloud Computing.

Ali Akbar Shah got his B.S. degree from Department of Computer Science and Information Technology in Iqra University in Pakistan in 2016 and doing his MS degree from Department of Computer Science and Technology in Shaanxi University of Science and Technology in China. His research interest is focused on Internet of Things, Android Operating System and Cloud Computing.

Design and Development of AI-based Mirror Neurons Agent towards Emotion and Empathy

Faisal Rehman[1]
Department of Computer Science
Government College University
Faisalabad, Pakistan

Awais Qasim[4]
Department of Computer Science
Government College University, Lahore, Pakistan
School of Science, Engineering and Environment
University of Salford, UK

Adeel Munawar[2], Aqsa Iftikhar[3], Jawad Hassan[5]
Fouzia Samiullah[6], Muhammad Basit Ali Gilani[7]
Department of Computer Science
Lahore Garrison University, Lahore, Pakistan

Neelam Qasim[8]
Lahore Business School
The University of Lahore, Lahore, Pakistan

***Abstract*—Since numerous years, researchers have to outline keen operators to accomplish the Artificial General Intelligence. Each new science revelation is an open challenge to all researchers. More than twenty years prior to a group of researchers discovered exceptional cerebrum cells, called reflect neurons in monkeys. These cells gave off an impression of being actuated both when the monkey accomplished something itself and when the monkey basically watched another monkey do a similar thing. This new discovery opened a new door for a scientist because of Mirror Neurons functionalities that can be huge contribute to cognitive science, neuroscience, impacting on Artificial General Intelligence. Mirror neuron functionality improves the Machine's learning. This research paper develops models for social interaction in which a machine may have the ability to learn the next person emotional state using mirror neurons and show empathy towards emotions.**

Keywords—Mirror neurons functionalities; emotions; empathy; machine learning; artificial intelligence

I. Introduction

For many years scientists are designing intelligent agents to achieve the target of Artificial General Intelligence. Hence every new scientific discovery is an open challenge to all scientists. More than twenty years, a scientist's team, at the University of Parma led by Giacomo Rizzolatti, find out special brain cells, which is called mirror neurons, in monkeys. Mirror neurons cells are activated when the monkey did anything itself and also when the monkey just watched another monkey do the same thing.

Different experiments have been done on the human brain using FMRI (Function Magnetic Resonance Imaging) that have shown that the human superior parietal lobe and inferior frontal lobe of human brain region neurons get activate when any action done by any person and also when any human experience another individual doing that same action. This has been recommended that these brain sections hold mirror neurons, so this is called the human mirror neuron system [1].

Mirror neurons have a direct communication link system between the sender and receiver of a message [2]. Mirror neurons mechanism develops very helpfully for understanding the message actions of one individual perform [3]. Some researchers believe that Mirror neurons have a link with Autism because Brain area that is having mirror properties experimented by EEG was less in children with autism [4].

Ramachandran was claimed that human self-awareness neurological basis was obtained from mirror neurons. Mirror neurons are not just to mirror outside but it can be helpful to know inward. In 2009 Ramachandran was written an essay for the Edge Foundation that provided the clarification theory for this “I additionally bet that these neurons can help emulate other individuals' conduct as well as can be turned 'internal' so to speak to make second-arrange portrayals or meta-portrayals of your own prior cerebrum forms. This could be the neural premise of contemplation, and of the correspondence of mindfulness and different mindfulness. There is clearly a chicken-or-egg question here as to which developed, to begin with, however... The fundamental point is that the two co-advanced, commonly enhancing each other to make the develop portrayal of self that describes current people" [5]. Behavior Recognition and Generation are basically referring when an individual observes another person's action, then deliberately performs that same action. Numerous specialists trust that programmed impersonation is intervened by the mirror neurons framework. Although automatic impersonation gets contribution by attentional procedures and yield by inhibitory procedures that is the reason it is long haul sensorimotor affiliation that can't be modified by deliberate procedures [6]. The combination of research on engine mimicry and programmed impersonation could uncover conceivable signs that these wonders rely upon the same mental and neural procedures [6] [7].

Nevertheless, due to the similarity of mirror neurons and automatic imitation, some researcher's need to suggest that programmed impersonation that is driven by the Mirror neurons framework. Programmed impersonation can be utilized as an apparatus to examine how the mirror neuron framework adds to subjective working and how engine mimicry advances mundanely demeanors and conduct [8] [9].

Many scientists have been performed experiments using different brain scanning techniques e.g. using FMRI (Functional Magnetic Resonance Imaging), EEG (Electroencephalography), anMEG (Magnetencephalography). These experiments have shown that brain regions are active when participants experiencing an emotion and when they see that same emotion experiencing by another person [10] [11] [12]. According to [13] that people who are empathic nature have strong activation in the mirror system for emotions.

In this paper, we proposed a model by which Machines can improve learning by mirror neuron functionality in which a machine may have the ability to learn the next person emotional state using mirror neurons and show empathy towards emotions. In this purposed model, human-machine interaction is created to test mirror functionalities. One human agent is interacting with a motivational agent. For instance, when any person is happy than another person mental state also gets changed due to mirror neurons. This can be for some seconds or last for more time depending on its intensity. Many experiments have been performed to know another person's mental state with respect to emotional conditions. Empathy is an ability to understand other's feelings as own emotions. The proposed Model agent must have the ability to learn the next person's emotional state by using mirror neurons through observation and experiences also Man-machine social interaction can predict the opponent's emotional state and respond with the same feelings depending on the intensity of emotional state.

Machines can improve learning by mirror neuron functionality. Therefore, there is a need to develop models for social interaction in which a machine may have the ability to learn the next person emotional state using mirror neurons and show empathy towards emotions.

II. Literature Review

This segment presents a brief introduction of the Mirror Neurons concept and structure. Human Brain performing Mirror Neurons functionality is also discussed. It reviews related work of Mirror Neurons in Neuro Science terms and cognitive Science. Proposed models for Mirror Neurons having different functionalities also discusses.

In the 1980s the neurophysiologist, Giacomo Rizzolatti with his colleagues was working on macaque monkeys to study neurons. The experiments were to allow the monkey to reach for a piece of food and meanwhile neurons were recorded. In this experiment, the researchers have been found that some neurons recorded while the monkey saw for a piece of food as well as when reach for that piece of food [14] [15].

The first experiment of Mirror neurons was carried on Macaque Monkeys. Mirror neurons are found in the inferior frontal gyrus (F5) and the inferior parietal lobe [16]. A recent experiment by Ferrari and his colleagues exposed that infant macaques can imitate a human face with a temporal period [17].

In the first experiment the monkey was just watching a piece of peanut, the pre-motor cortical cells that have been active when the monkey was reaching toward the piece of peanut. In Fig. 1 shows that same brain area gets active when monkey observing someone doing the same activity. A Strong activation is present in F5 during observation of the experimenter's grasping movements, and while the same action is performed by the monkey [18].

Different experiments have been done on the human brain using functional magnetic resonance imaging (fMRI) that have shown that the human superior parietal lobe and inferior frontal lobe of human brain region neurons get active when the person does an action and also when the person experiences another individual doing that same action. It has been suggested that these brain regions contain mirror neurons, and they have been defined as the human mirror neuron system as shown in Fig. 2 [1].

Many experiments have been done on the Human Brain to know the Mirror neurons functionality. Their experiments have been done by functional magnetic resonance imaging (fMRI), EEG and MEG.

According to cortical homunculus that is a representation of the functional divisions of the primary motor cortex of the human brain that is directly responsible for the motor information of body movement and same as the primary somatosensory cortex that is directly responsible for the movement and exchange of sensory information of the body in the human brain. Brain activation in frontal and parietal areas during the observation of mouth, hand and foot actions.

The experiment held by [19] on 14 healthy right-handed volunteers. This experiment using fMRI found that the secondary cortex is activated when the participants observe someone or someone getting touched by some object.

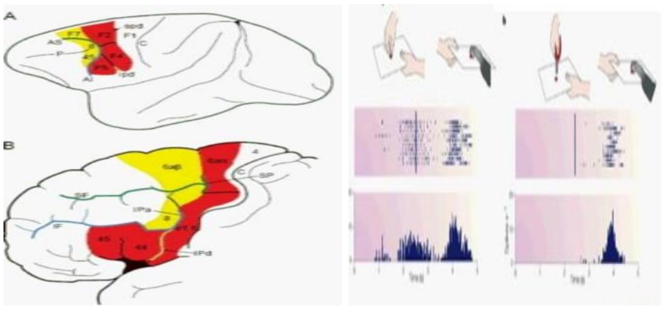


Fig. 1. Monkey Brain Get Activated while Observing and Reaching to a Piece of Peanut [18].

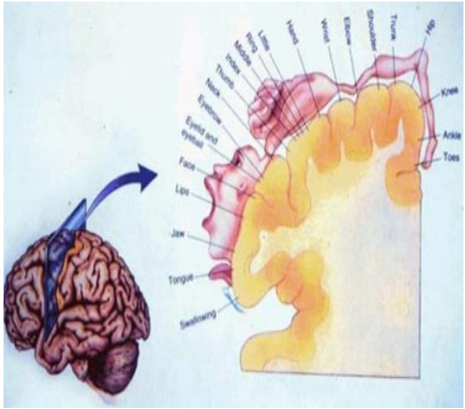


Fig. 2. Human Brain and Mirror Neurons.

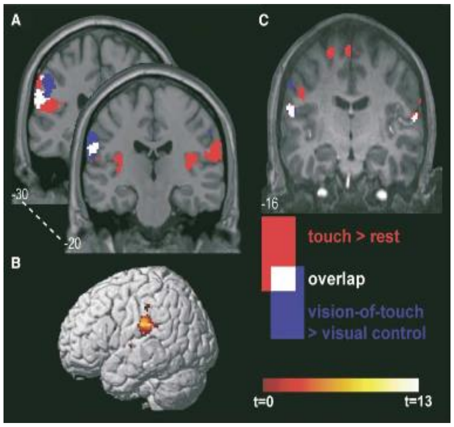


Fig. 3. Experiment Result using fMRI.

Fig. 3 illustrates the extent of the over-lap between touch and vision-of-touch. Brain area is activated in red color is showing when someone is touched. Blue areas of the brain get activated when vision-of-touched performed. While white color indicated overlap of these both actions.

Mirror Neurons in Neuro Science an adaptive agent model was proposed that provides an evolutionary link between imitation and Mirror Neurons [20].

The recent research is conducting by a researcher that's the focus is the visual recognition of goal-directed movements. The basic idea is to understand the intentions and action goals of others that can be possible by Mirror neurons. [21].

A neural and cognitive Model was proposed that gives an abstract neural model that is mapped on the cognitive level model [22].

The emotional Module maintained an emotional state of a machine. This module had a bidirectional link with the drive and behavior module. A behavior module was used to take action in order to meet the aim. Simulation of Glucose and Insulin Theories for the Implementation of Psychophysiology Drive Regulatory System in QuBIC Agents model is proposed by [23].

Many Researchers have been argued that Mirror Neurons are involved in Empathy. Many scientists have been performed experiments using different brain scanning methods e.g. using fMRI (Functional Magnetic Resonance imaging), EEG (Electroencephalography), and MEG. These experiments have shown that brain regions are active when participants experiencing an emotion and when they see that same emotion experiencing by another person. [10] [11] [12]. According to [13] that people who are empathic nature have strong activation in the mirror system for emotions. Mirror neuron's functionality is being performed by using emotions and empathy behavior.

Empathy is an ability to understand other's feelings as own emotions. In Fig. 4 an experiment was held to know participants' neural basis for understanding others' emotions. In these experiments, fMRI (Functional Magnetic Resonance Imaging) is used that scanned the human brain at once to know about the activated areas of the brain with respect to current activity [24].

Fig. 4. Experiment Results [24].

Six basic emotions were proposed by [25] which is also known as Universal emotions. Emotion extraction from NLP has significant importance in Artificial Intelligence. Emotions have been having importance in psychological and behavioral sciences. There are different types of approaches used for emotion detection. This can be categorized mainly in keyword-based approaches, linguistic rules-based and can be machine learning techniques. The keyword-based approach can be applied to simple models because it cannot handle all the cases [26].

However, linguistic rules-based is computational linguistics rules that define language structure. In this regard, ESNA system was developed to classify news headlines [27]. The latest rule-based approach can recognize nine emotions [28]. Another approach is used in linguistic rule-based was metaphorical data that was more practical as there was any set of emotions [29].

The machine learning approach is based on statistical techniques that can be further divided into supervised learning or unsupervised techniques. A large amount of information is required to train data sets in supervised learning. Support Vector Machines have been used to classify different blog sentences [30]. One research is conducted to compare three machine learning algorithms that concluded that Support Vector Machine performance was best [31]. While unsupervised learning is another approach; using these techniques 'LSA Single word' is proposed that calculates the similarity between texts. This approach was using WordNet synsets [32].

III. Mirror Neuron Functionality in Motivational Architecture

This segment presents a model extension to implement mirror neurons functionality using its emotional empathy concept. It discusses different existing modules of the model and also introducing new modules as an extension of the model.

A. Mirror Neuron Motivational Model Extension

This model which is shown in Fig. 5 is an extension of Simulation of Glucose and Insulin Theories for the Implementation of Psychophysiology Drive Regulatory System in the QuBIC Agents model is proposed by (Khan) discussed in chapter two. This model has been extended by external input from the environment to achieve the mirror neurons functionality by communicating the existing agent.

B. Mirror Neurons- Emotions and Empathy

In this architecture, human-machine interaction is created to test mirror functionalities. One human agent is interacting with a motivational agent. For instance, when any person is happy then another person's mental state also gets changed due to mirror neurons. This can be for some seconds or last for more time depending on its intensity. Many experiments have been performed to know another person's mental state with respect to emotional conditions. Empathy is an ability to understand other’s feelings as own emotions.

1) Environment: According to human- machine-based interaction, there is an environment that creates a link between humans and machines. This link provides interaction between humans and machines to act accordingly. A Human from the environment can have interacted with the machine agent through this environment. The environment is also responsible to define a rule for any social interaction.

Emotion is an aspect of a person’s mental state that is a person’s internal (physical) and external (social) sensory feeling [33]. This architecture is providing some input from the environment and its effect on internal drives of physical states. This is word-based for emotion calculations. There is input which is a sentence or a collection of words. This information is calculated on measures to find proper emotions that hidden in words and expressions.

2) Universal emotions: As discussed in Section 2, six basic emotions were proposed by (Ekman) which was Happiness, Sadness, Fear, Disgust, and Anger. Four out of these six emotions are negative while just two are positive. This information was gathered from different cultures. This research revealed that there is a Universal Set of Emotions categorized in six [25].

WordNet effect is a list of words that are categorized into six basic emotions which were called the Universal Set of Emotions [34]. In this architecture, the WordNet Affect list is being used to identify emotions from the text. Emotions can be extracted easily from facial expressions and even from voice data. Emotions extraction from text is something different. The same words can have different senses. In this architecture, emotion is recognized from text input. Firstly text is analyzed and tagged with emotionally identified words with intensity.

3) Sentence analysis: Input can be a word or combination of words in the form of a sentence. There can be different kinds of sentences, Interrogative sentences, negative sentences, declarative sentences, or a simple sentence. The purpose of this research is to simulate the mirror neuron's functionality and create a Human-Machine interaction. The input sentence is information in the text that will be a break in words, there are two types of the corpus that is helping in identifying emotions in the text.

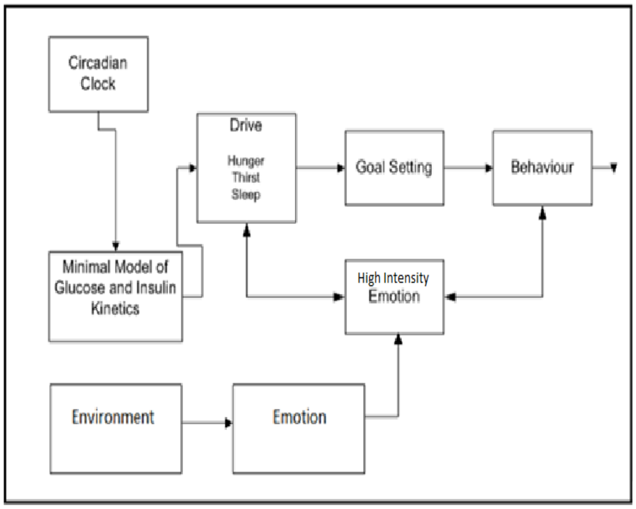


Fig. 5. Mirror Neurons Extension with Motivational Architecture.

WordNet Affect list is used as a built-in corpus for emotional recognition. Another corpus is implemented based on context analysis of the text. The same sentence can vary in a sense. To make the system work efficiently, it is necessary to make its context analysis. Sometimes sentence means a normal feeling but to what person is said matters. Human interacts with each other according to relations and social environment. Mirror neuron's emotions can have different intensity levels depending upon relationships with the next person.

For instance, if we see some person in pain, then it is obvious our mirror neurons will active our painful emotions. While with friends we mimic their activities and make fun of each other by mirror neurons functionality, but this same attitude cannot be followed for respected level persons. So it is very important to know the next person context while analyzing sentences. In this architecture, there is an agent who is regulating its drives and goals to perform some specific behavior against goal achieved. Therefore, a context is related to this agent to know and evaluate its behavior that can change drives.

The input sentence is providing an emotional response that is changing agent emotional state with respect to its intensity level. This changing emotional state is having a link with agent internal drives such as Glucose, insulin, etc.

4) Emotion extraction: In this architecture, six basic emotions are targeted which has been discussed in perivious section. These emotions are also known as Universal emotions. There are two positive and four negative emotions. Emotion can be recognized by facial expressions, from voice analysis and text. In this architecture, text emotions are used to simulate mirror neuron’s behavior.

Fig. 6 shows the text of input from the environment section module. This module is basically implemented to achieve mirror neuron’s functionality. An agent or human from the environment is interacting with the existing system agent to know its mirror behavior towards drives and goals. This emotion can change the next agent’s feelings with respect to its intensity level. The model agent has own emotional state;

which is regulating its drives and goals to perform specific functions. Meanwhile, environment interaction with the model agent can change the whole scenario if the environment emotion intensity level is high than model emotions intensity. High-intensity emotion can change internal drives which are Hunger, Thirst, and Sleep. Emotions are also attached to physiological parameters. These changes can change the goals and behavior of a model agent accordingly.

This task is achieved by a Human-Machine interaction to know the mirror functionality. An environment agent talks with the model agent in text form. This text is the basic information that will change model agent behavior to know mirror functionality. The process flow of this module will be as follows.

Fig. 7 shows process flow of emotion extraction from text. There is an input text from the environment that can be sentenced. In the first step, the sentence will be analyses contextually. Contextual analysis has significant importance in the Mirror Neurons functionality. Because while performing Mirror activity; it cannot be the same with all the agents. For instance; the behavior of a student with teachers and friends will be change. In the case of mimic as major functionality of mirror neurons can depend on a relationship with the next person. While emotions and empathy is a sub functionality of Mirror neurons that can depend on intensity level with respect to the relationship with the next person. Sometimes while going on the road, if we saw anyone in pain by having some accident; then this can affect our internal emotional feelings but the time period of this feeling can be varied. However, if we see this same situation with our close one then it can be for a longer time period.Further, the sentence will be a break in words; as discussed in chapter two that WordNet effect is a list of words that are categorized into six basic emotions. This list provides help in identifying words with a direct emotional category. Some words have high-intensity levels while some have normal or medium. This process will tag words with emotional feelings. After this, both level intensity is analyses and assigned one final emotion.

Now the model agent is having own emotional feelings, and an environment Human interacts with another emotional feeling. This is the main step to finalize the mirror functionality. Now it is depending on the intensity level of the model agent emotion and environment emotion. For example, if the model agent is feeling hungry with high intensity of negative emotion; and environment emotional intensity is low or medium; this can affect model agent physiological drives with slight increase or decline, but high-intensity feelings will be performed first.

High-intensity emotion will be set that will affect on physiological parameters. These parameters can change the model agent drives. There are three basic drives, Hunger, Thirst or Sleep. These drives can set goals specifically and the behavior of the model can be changed accordingly. Each drive has its own level which is depending on physiological parameters. For instance, the glucose level is low then the model agent starts feeling hungry. Similarly, the volume parameter is linked with feeling thirst, a low level of volume can arise thirst feeling; while after drinking water this level gets maintained and the model agent got happy.

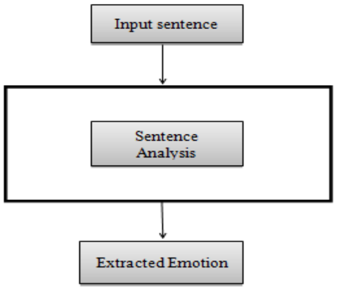


Fig. 6. Abstract view of Emotion Extraction.

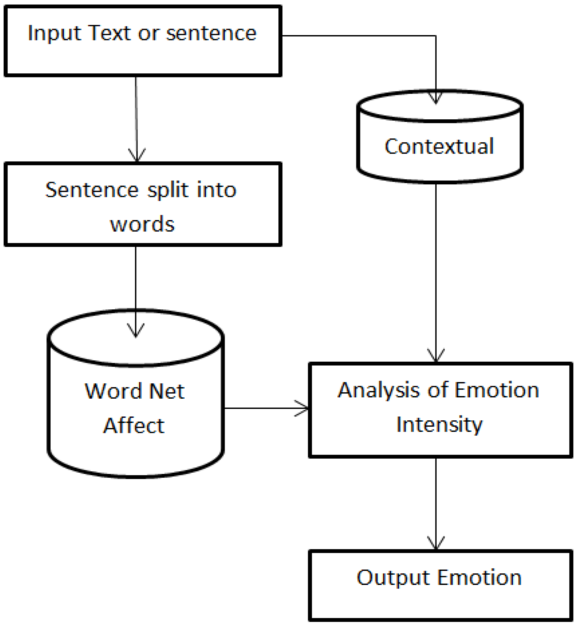


Fig. 7. Process flow of Emotion Extraction from Text.

C. Motivational Architecture

Several modules such as emotional state, biological clock, level of glucose or insulin, target orientation or attitude module work in a system. Each of them has its own role to perform when it comes to generating inspiration in a system. The motivational model functions in a circadian clock in order to maintain a certain level of drive in a system. The human body thus works according to a clock that is linked up to the solar system and hence a person craves or sleeps accordingly. Such activity ends up maintaining a certain level of glucose and insulin.

A stable range of insulin/glucose level, the temperature state, sodium level and its volume, concentration of the solution, etc are few of the parameter which drives module to manage so that the hunger, thirst and sleep works at an optimum level. Setting a module set is most important to monitor the drive state.

1) Circadian clock: The circadian clock is an internal biological clock that helps to organize internal and external activities of the human body shown in Fig. 8. It's a 24-hour cycle that regulates basic human drives and functions and used in motivational architecture to generate motivation in a continuous manner a clock was needed to simulate a natural drive regulatory system in machines. The circadian clock is running in the background of our brain and create a cycle of sleepiness and alertness at regular intervals. The circadian clock is utilizing as a clock that ticked on consistent premise to motivate machine. It is controlled by the release of hormones that are regulated b the brain. It operated on a minimal model module to get levels of insulin and glucose continuously.

2) Drive module: The drive module is supposed to basically support three physiological needs i.e. Hunger, Thirst and Sleep. It keeps a proper check at every measure that can have an influence over any of the above-mentioned drive. The drive status can be noted in numeric range and hence that range should be maintained stable or else the instability of the drive state can lead to many health issues. The insignificant model module helps to measure the required parameters. The parameters tend to change as shown by the insignificant model. The level of glucose and insulin has a strong impact on physiological needs (Hunger, Thirst, and Sleep). Thus, it has been concluded that the level of glucose and insulin are the two main factors that stimulate the feelings of being hungry, thirsty or sleep.

3) Goal setting module: There are three main goals in a motivational machine:

- Hunger
- Thirst
- Sleep

Drive strengths were compared and a drive with the high strength was selected as a goal. For Example, if thirst has the highest drive value other than remaining elements, then all the other drivers will search for water, in the order to quench the thirst.

The goal-setting module is shown in Fig. 9 by an Artificial Neural Network (ANN). Inputs are received from the Drive model and calculated from all physiological parameters on the basis of glucose and insulin. The input of the goals setting model is received from six physiological parameters such as Osmolarity, Glucose, Volume, Sodium, Insulin, and Temperature.

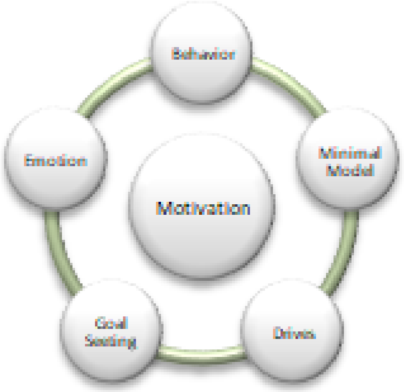
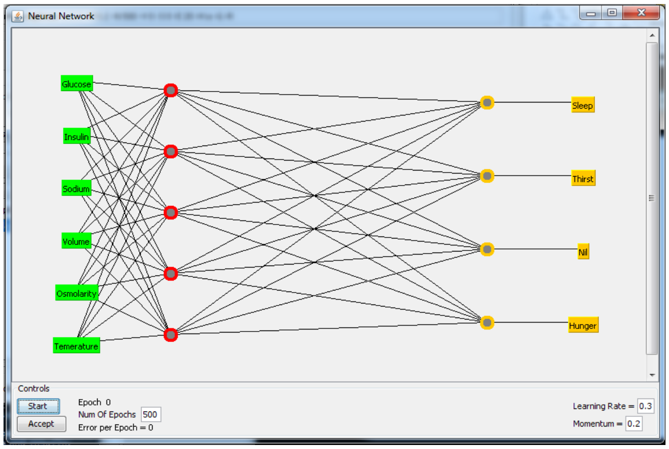


Fig. 8. Motivational Architecture Circadian Clock.



Epoch 0
Num Of Epochs 500
Error per Epoch = 0
Learning Rate = 0.3
Momentum = 0.2

Fig. 9. Goal Setting Module Neural Network.

4) Emotion module: Emotions are derived from feelings and are divided into two categories:

- Negative Emotions
- Positive Emotions

The current emotional state is delivered by Drive state. There is a bidirectional link with the Behavior module and drive. Drive module helps to strengthen different drives based on physiological parameters. These strengths can be low, medium or high. High and low strength presents and sets a negative state and while medium-strength presents a positive state. A low value can create a negative emotional state which can make a machine angry while balanced glucose set a positive emotional state which makes a machine happy. As discussed above along with drive state emotional is also set by Behavioral module. The behavior module is responsible for taking action after the goal has been set by goal setting module.

The Artificial Neural Network (ANN) of the emotional module is given in Fig. 10. This module takes input from the drive module and behavior module that is in physiological parameters. This parameter strength used to select appropriate emotion.

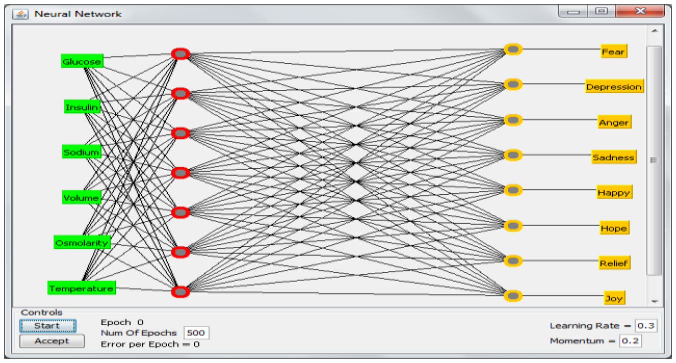
5) Behaviour module: The basic function of this module is to relate to any feature that is most desirable to achieve a target. The behavior module works in an imitated environment. For further elaboration an experiment is been quoted here- Behavior module is tested in a path puzzle called a maze. With several cracks down along the path, some food and water were placed in a maze. On three different positions, food and water were placed with different levels of glucose and sodium in it. The subject (who has to get to the food in a maze) wanders here and there in complete hustle before he decided on which position to go to get the food. Hence proven, the goal is required to have a motivation. The subject showed a few kinds of behaviors for the purpose of getting food, water and go for rest.

D. Working of Architecture

This architecture is recording the feeling and response of a person according to the release of neurons in his body. The human psychological parameters tend to maintain his three basic drives. The biological clock that works with the sunrise and sunset supports the need drive of the human body. On each specific time, a specific range of glucose and insulin is being produced by the body on a daily basis. The minimal model helps to note down the value of the level of glucose and insulin. The goal-setting module received the values and strength of each drive been calculated by the drive module. The emotional module analysis the strength of drive and sets the emotional statistics, for instance, if the strength is in normal range then the emotional state will be positive and when strength is in high range the emotional state will be negative. The emotional state of a system influences his three basic drives. The level of the drive strength varies e.g. when a person is mad at something, he might feel hungry or if when he is pleased, he doesn't care much about his food. Drive strength works for goal setting module. This module already had a regular range for each drive. The behavior module stimulates the actions that are needed to achieve a target. The human interaction can change the drive state of the subject and can also lead him to a different course of action.

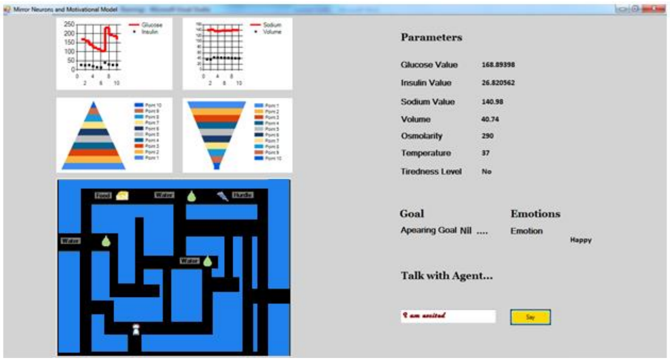
E. Application View

This section displays variants states of Mirror Neuron functionality in screenshots with Motivational Model application. This section also shows the screenshots of the application when this model is communicated by environmental agents that can change existing goals such as hungry, thirsty and sleepy depending on physiological parameters change.



Epoch 0
Num Of Epochs 500
Error per Epoch = 0
Learning Rate = 0.3
Momentum = 0.2

Fig. 10. Neural Network of Emotional Module.



Parameters

Glucose Value 168.89398
Insulin Value 26.820562
Sodium Value 140.98
Volume 40.74
Osmolarity 290
Temperature 37
Tiredness Level No

Goal
Apearing Goal Nil ....

Emotions
Emotion Happy

Talk with Agent...

I am excited

Say

Fig. 11. Mirror Neurons and Motivational Architecture-Application view.

Fig. 11 shows the view of the application. The application is divided into different sections. There are physiological parameters that are changing with respect to model agent movements. These internal parameters regulate internal drives and set goals that activate the behavior of the agent. Behaviour is a specific action against the targeted goal.

Drives with moderate strength set positive emotions on the other hand drivers with intense strength set negative emotions.

The results were as below:

- The emotional state was also affected by Hunger which created different feelings such as anger, distress, anxiety, and sadness.
- Thirst also changed the emotional state and created negative emotions.
- Lack of sleep created anxiety, sadness, and fear [35-43].

It has been argued in [44-46] that the use of intelligent agents for the implementation of intelligent systems is highly desirable. The main objective of this Actions Selection Module was to achieve the goal by adopting a behavior.

IV. Results and Discussion

Experiments have been performed to simulate the Mirror Neurons functionality and its effect on motivational architecture. Some experiments have been stated here showing the results of different modules. These experiments are discussed in changing all the parameters and their effects on existing states. Changing behavior is also observed and discussed with results.

A. Experiment 1

1) Analysis of experiment 1: Table I shows a simple scenario when the system is having nil goals and feeling is also having a medium intensity level. Agent existing mode of goal and behavior plays an important role in behavior. Goal and behavior can be changed by the intensity level of emotion. This intensity can be high, medium or low depending upon the current state of the agent.

B. Experiment 2

1) Analysis of experiment 2: Experiment 2 results are shown in Table II. This experiment is observed to know the values of parameters without environment agent communication. In this experiment verify the supposition and literature made in this research. The experiment exposed the results whenever the normal condition of physiological parameters and homeostatic range maintained then expose the moderate behavior. During normal conditions it showoff the emotion of happiness, joy, and relief with positive emotions. It was also proved with experiment results in Table II in which the homeostatic range model did not set any certain goal. According to drive priority, mostly occurring drive was thirst and felt before any primary drive especially before sleep and hunger. This model feels thirst before any other drive, it can be seen in experiments.

TABLE. I. Emotion and Goal Experiment

Sentences	Emotion	Goals
I was excited when was opening my gift.	Surprise	Nil
Some students like to study in the morning	Happiness	Nil
She was sad because of low grades in exams.	Sadness	Nil
I like to hang out with my friends or family.	Happiness	Nil

TABLE. II. Experiment Results

Glucose	Insulin	Sodium	Volume	Osmolarity	Temperature	Tiredness	Emotion	Goal
170.89	23.47	139.87	37.77	290	37.2	No	Positive	Nil
155.10	22.22	140.11	37.48	290	37.2	No	Positive	Nil
142.23	19.75	141.17	37.14	290	37.2	No	Positive	Nil
117.15	19.09	142.28	36.55	290	37	No	Positive	Nil
111.67	12.37	142.47	36.10	290	37	Low	Positive	Nil
101.35	11.57	143.05	36.02	291	37.2	Medium	Negative	Thirst
After Satiation of Thirst								
95.23	7.25	135.27	44.2	290	37.2	No	Positive	Nil
89.145	7.01	135.88	43.75	290	37.3	No	Positive	Nil
87.34	6.55	136.16	42.50	290	37	No	Negative	Hunger
After Satiation of Hunger								
350.37	94.2	137.11	41.92	290	38	No	Positive	Sleep
After Waking Up								
190.80	29.019	138.66	41.87	290	37	No	Positive	Nil
188.05	27.20	140.40	41.51	290	37	No	Positive	Nil
168.05	26.68	140.94	41.13	290	37	No	Positive	Nil

C. *Experiment 3*

This experiment is implemented to know the changing behavior of the system by environment communication. This experiment is giving mirror neurons functionality. The external text input can change the behavior, feelings, and goals for some specific time period. In these experiment shows examples with results and variation in internal parameters.

1) Mirror neuron and sleeping state: Table III shows result for Mirror Neuron and sleeping state. A high level of glucose causes sleep. And glucose increased by taking food. After fulfilling hunger glucose level got high that is the reason for feeling sleepy. The model started feeling sleep when glucose level was above 300 mgdl. While the sleeping agent is communicated by the external environment with negative emotion. The intensity was high of external emotion so the model agent felt empathy towards and came in sadness feelings. This change of feelings causes of changing all the physiological internal parameters. These parameters set goals and behavior with respect to the current situation.

2) Mirror neuron and hunger state: Table IV shows result for mirror neurons effect on hunger state. Glucose level maintains the food level. With the decline in glucose and insulin, the hunger level gets activated with anger emotion. Even the consumption of water failed to raise the level of glucose and insulin. The values of glucose and insulin were down at the time thirst but due to high priority of thirst model set thirst as a goal. Glucose theory says that a low level of glucose makes us hungry. Results show that when glucose was its low level of 87.34 mgdl, the model started feeling hungry.

Glucose has a positive relation with insulin, high glucose raises the level of insulin and vice versa. That’s why the huger level of insulin was also at its lowest point of 6.55. The experiment confirmed all theories. It also showed that during the intense sensation of hunger emotional state of hunger becomes negative. After the fulfillment of desire glucose and insulin get back to their normal range and the Emotional state also becomes positive.

When the model was in a hunger state with a low level of glucose and insulin; then an agent is communicated with the architecture agent to simulate its mirror functionality. The external agent from the environment gives input text which was positive. After taking positive influence from environment physiological parameters values got changed. However, internal feelings were having high intensity so the emotion remained negative but the mood of the architecture agent is enhanced. Mirror neurons can have a significant impact on human internal drives that can change internal physiological parameters even though when drives are activated in high-level demand.

3) Mirror neuron and thirst state: Table V shows the result for mirror neurons and thirst state. The experiment suggested that before the environment interaction model started feeling thirsty. Thirst feelings became passionate when osmolarity got high and greater than 290 mmol/kg. Osmolarity defined the number of osmoles (concentration of salts) per kg. Table V showed that osmolarity was 291 mmol/kg at the time of thirst. Thirst was also dependent on the level of sodium; sodium kept on increasing with time, thirst was felt when the concentration of sodium level was higher than 142 mEq/l.

Experiment result presented that while at a state of thirst sodium level was 143.6 mEq/l; which proved that higher concentration salts caused thirst. The volume of water was another factor that affected thirst. A low volume of water created thirst which was already proved in the experiment table. The result also shows that intense drives created negative emotions. The Parameters get back into their homeostatic range when the thirst was quenched. The volume of water also went up to 43.38, Osmolarity became 290 and the level of sodium also decreased to 135.66.

The model was thirsty when an agent from the environment interacted with the model in a positive way. Due to these interactions mirror neurons get activated that changed the internal physiological parameters. Osmolarity was on its same level 290 mmol/kg, which means that that interaction just helped to change the physiological parameters from negative to positive but the model remains still thirsty. The demand for water was reduced by stabling the Sodium level by 141.62 mEq/1 and emotion become positive. Mirror Neuron had a great impact on physiological parameters that helps to normalize internal drives.

TABLE. III. MIRROR NEURONS EFFECT ON SLEEPING STATE

Glucose	**Insulin**	**Sodium**	**Volume**	**Osmolarity**	**Temperature**	**Tiredness**	**Emotion**	**Goal**
350.37	94.2	137.11	41.92	290	38	No	Positive	Sleep
Parameters after Human interaction with an agent with the following input sentence								
Input sentence: She was sad because of low grades in exams.								
168.89	26.82	140.98	40.74	290	37	No	Negative	Nil

TABLE. IV. MIRROR NEURONS EFFECT ON HUNGER STATE

Glucose	**Insulin**	**Sodium**	**Volume**	**Osmolarity**	**Temperature**	**Tiredness**	**Emotion**	**Goal**
95.23	7.25	135.27	44.2	290	37.2	No	Positive	Nil
89.145	7.01	135.88	43.75	290	37.3	No	Positive	Nil
87.34	6.55	136.16	42.50	290	37	No	Negative	Hunger
Parameters after Human interaction with agent with following input sentence								
Input sentence: Why you are down? You can do it.								
102.37	12.57	138.64	40.74	42.32	37	No	Negative	Nil

TABLE. V. MIRROR NEURONS AND THIRST STATE

Glucose	**Insulin**	**Sodium**	**Volume**	**Osmolarity**	**Temperature**	**Tiredness**	**Emotion**	**Goal**
206.42	36.24	137.66	39.58	290	37.7	No	Positive	Nil
200.56	33.95	138.32	39.16	290	37.7	No	Positive	Nil
191.86	30.01	138.98	38.74	290	37.2	No	Positive	Nil
188.15	29.29	139.64	38.32	290	37.2	No	Positive	Nil
168.89	20.82	140.3	39.90	290	37.2	No	Positive	Nil
160.01	25.22	140.96	37.48	290	37.2	No	Positive	Nil
139.74	20.25	141.62	37.06	290	37.2	No	Positive	Nil
122.16	19.76	142.28	36.64	290	37	No	Positive	Nil
113.69	14.33	142.94	36.22	290	37	Low	Positive	Nil
102.37	12.57	143.6	35.8	291	37.2	Medium	Negative	Thirst
Parameters after Human interaction with an agent with the following input sentence								
Input sentence: I was excited when was opening my gift.								
160.03	25.22	141.62	37.06	291	37	Medium	Positive	Nil

4) Analysis of experiment 3: The results of experiment 3 showed that Mirror neurons functionality has a strong impact on physiological parameters. These changes regulate the internal drives; that effect in goal setting of the model and behavior can be changed. Feelings can be changed to positive with environment changes.

V. CONCLUSION AND FUTURE WORK

This paper represented the Mirror Neurons functionality which has different aspects and it also focused on emotions and empathy. Motivational architecture is used to validate its functionality of mirror neurons. In the next version of this architecture can be improved by making more generic interaction with the agent. This paper helps the machine interaction with human and also provide an architecture design to learn from human by getting meaningful information from a sentence. This model is initially designed for the English language, it can be enhanced for more languages like Urdu, Arabic, Turkish and Hindi in the future.

REFERENCES

[1] Iacoboni, Marco, Roger P. Woods, Marcel Brass, Harold Bekkering, John C. Mazziotta, and Giacomo Rizzolatti. "Cortical mechanisms of human imitation." science 286, no. 5449 (1999): 2526-2528.

[2] Rizzolatti, Giacomo, and Michael A. Arbib. "Language within our grasp." Trends in neurosciences 21, no. 5 (1998): 188-194.

[3] Rizzolatti, Giacomo, and Maddalena Fabbri-Destro. "The mirror system and its role in social cognition." Current opinion in neurobiology 18, no. 2 (2008): 179-184.

[4] Oberman, Lindsay M., Edward M. Hubbard, Joseph P. McCleery, Eric L. Altschuler, Vilayanur S. Ramachandran, and Jaime A. Pineda. "EEG evidence for mirror neuron dysfunction in autism spectrum disorders." Cognitive brain research 24, no. 2 (2005): 190-198.

[5] Oberman, Lindsay M., and Vilayanur S. Ramachandran. "Reflections on the mirror neuron system: Their evolutionary functions beyond motor representation." In Mirror neuron systems, pp. 39-59. Humana Press, 2008.

[6] Longo, Matthew R., Adam Kosobud, and Bennett I. Bertenthal. "Automatic imitation of biomechanically possible and impossible actions: Effects of priming movements versus goals." Journal of Experimental Psychology: Human Perception and Performance 34, no. 2 (2008): 489.

[7] Van Baaren, Rick B., William W. Maddux, Tanya L. Chartrand, Cris De Bouter, and Ad Van Knippenberg. "It takes two to mimic: behavioral consequences of self-construals." Journal of personality and social psychology 84, no. 5 (2003): 1093.

[8] Heyes, Cecilia. "Automatic imitation." Psychological bulletin 137, no. 3 (2011): 463.

[9] Paukner, Annika, Stephen J. Suomi, Elisabetta Visalberghi, and Pier F. Ferrari. "Capuchin monkeys display affiliation toward humans who imitate them." Science 325, no. 5942 (2009): 880-883.

[10] Botvinick, Matthew, Amishi P. Jha, Lauren M. Bylsma, Sara A. Fabian, Patricia E. Solomon, and Kenneth M. Prkachin. "Viewing facial expressions of pain engages cortical areas involved in the direct experience of pain." Neuroimage 25, no. 1 (2005): 312-319.

[11] Cheng, Yawei, Chia-Yen Yang, Ching-Po Lin, Po-Lei Lee, and Jean Decety. "The perception of pain in others suppresses somatosensory oscillations: a magnetoencephalography study." Neuroimage 40, no. 4 (2008): 1833-1840.

[12] Jabbi, Mbemba, Marte Swart, and Christian Keysers. "Empathy for positive and negative emotions in the gustatory cortex." Neuroimage 34, no. 4 (2007): 1744-1753.

[13] Gazzola, Valeria, Lisa Aziz-Zadeh, and Christian Keysers. "Empathy and the somatotopic auditory mirror system in humans." Current biology 16, no. 18 (2006): 1824-1829.

[14] Di Pellegrino, Giuseppe, Luciano Fadiga, Leonardo Fogassi, Vittorio Gallese, and Giacomo Rizzolatti. "Understanding motor events: a neurophysiological study." Experimental brain research 91, no. 1 (1992): 176-180.

[15] Rizzolatti, Giacomo, Luciano Fadiga, Vittorio Gallese, and Leonardo Fogassi. "Premotor cortex and the recognition of motor actions." Cognitive brain research 3, no. 2 (1996): 131-141.

[16] Rizzolatti, Giacomo, and Laila Craighero. "The mirror-neuron system." Annu. Rev. Neurosci. 27 (2004): 169-192.

[17] Ferrari, Pier F., Elisabetta Visalberghi, Annika Paukner, Leonardo Fogassi, Angela Ruggiero, and Stephen J. Suomi. "Neonatal imitation in rhesus macaques." PLoS biology 4, no. 9 (2006).

[18] Pollick, Frank E. "Humanoids 2001 Tutorial on the Visual Perception of Human Movement."

[19] Keysers, Christian, Bruno Wicker, Valeria Gazzola, Jean-Luc Anton, Leonardo Fogassi, and Vittorio Gallese. "A touching sight: SII/PV activation during the observation and experience of touch." Neuron 42, no. 2 (2004): 335-346.

[20] Borenstein, Elhanan, and Eytan Ruppin. "The evolutionary link between mirror neurons and imitation: An evolutionary adaptive agents model." Behavioral and Brain Sciences 28, no. 2 (2005): 127-128.

[21] Oztop, Erhan, Mitsuo Kawato, and Michael Arbib. "Mirror neurons and imitation: A computationally guided review." Neural networks 19, no. 3 (2006): 254-271.

[22] Bosse, Tibor, Zulfiqar A. Memon, and Jan Treur. "A cognitive and neural model for adaptive emotion reading by mirroring preparation states and Hebbian learning." Cognitive Systems Research 13, no. 1 (2012): 39-58.

[23] Khan, Faiz M., and John P. Gibbons. Khan's the physics of radiation therapy. Lippincott Williams & Wilkins, 2014.

[24] Wicker, Bruno, Christian Keysers, Jane Plailly, Jean-Pierre Royet, Vittorio Gallese, and Giacomo Rizzolatti. "Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust." Neuron 40, no. 3 (2003): 655-664.

[25] Matsumoto, David. "Cultural influences on the perception of emotion." Journal of Cross-Cultural Psychology 20, no. 1 (1989): 92-105.

[26] Olveres, Jimena, Mark Billinghurst, Jesus Savage, and Alistair Holden. "Intelligent, expressive avatars." In Proceedings of the First Workshop on Embodied Conversational Characters, pp. 47-55. 1998.

[27] Al Masum, Shaikh Mostafa, Helmut Prendinger, and Mitsuru Ishizuka. "Emotion sensitive news agent: An approach towards user centric emotion sensing from the news." In IEEE/WIC/ACM International Conference on Web Intelligence (WI'07), pp. 614-620. IEEE, 2007.

[28] Neviarouskaya, Alena, Helmut Prendinger, and Mitsuru Ishizuka. "Recognition of affect, judgment, and appreciation in text." In Proceedings of the 23rd international conference on computational linguistics, pp. 806-814. Association for Computational Linguistics, 2010.

[29] Neuman, Yair, Gabi Kedma, Yohai Cohen, and Ophir Nave. "Using web-intelligence for excavating the emerging meaning of target-concepts." In 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 22-25. IEEE, 2010.

[30] Aman, Saima, and Stan Szpakowicz. "Using roget’s thesaurus for fine-grained emotion recognition." In Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I. 2008.

[31] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 79-86. Association for Computational Linguistics, 2002.

[32] Strapparava, Carlo, and Rada Mihalcea. "Learning to identify emotions in text." In Proceedings of the 2008 ACM symposium on Applied computing, pp. 1556-1560. 2008.

[33] Zhang, Yu, Zhuoming Li, Fuji Ren, and Shingo Kuroiwa. "A preliminary research of Chinese emotion classification model." Advances in Artificial Intelligence (2008): 95.

[34] Strapparava, Carlo, and Alessandro Valitutti. "Wordnet affect: an affective extension of wordnet." In Lrec, vol. 4, no. 1083-1086, p. 40. 2004.

[35] Macht, Michael. "How emotions affect eating: a five-way model." Appetite 50, no. 1 (2008): 1-11.

[36] Izard, Carroll E. Human emotions. Springer Science & Business Media, 2013.

[37] Izard, C. E. "Emotions, personality, and psychotherapy. The psychology of emotions. New York, NY, US." (1991).

[38] Jain, Dreama. "A Comparative Framework for Emotion Driven Agent Based Modelling." (2011).

[39] Baglioni, Chiara, Kai Spiegelhalder, Caterina Lombardo, and Dieter Riemann. "Sleep and emotions: a focus on insomnia." Sleep medicine reviews 14, no. 4 (2010): 227-238.

[40] Scott, Brent A., and Timothy A. Judge. "Insomnia, emotions, and job satisfaction: A multilevel study." Journal of Management 32, no. 5 (2006): 622-645.

[41] Thomsen, Dorthe Kirkegaard, Mimi Yung Mehlsen, Søren Christensen, and Robert Zachariae. "Rumination—relationship with negative mood and sleep quality." Personality and Individual Differences 34, no. 7 (2003): 1293-1301.

[42] Uhde, Thomas W., and Bernadette M. Cortese. "Anxiety and insomnia." In Anxiety in Health Behaviors and Physical Illness, pp. 105-127. Springer, New York, NY, 2008.

[43] Konidaris, George, and Andrew Barto. "An adaptive robot motivational system." In International Conference on Simulation of Adaptive Behavior, pp. 346-356. Springer, Berlin, Heidelberg, 2006.

[44] Qasim, Awais, and Syed Asad Raza Kazmi. "MAPE-K interfaces for formal modeling of real-time self-adaptive multi-agent systems." IEEE Access 4 (2016): 4946-4958.

[45] Qasim, Awais, Syed Asad Raza Kazmi, and Ilyas Fakhir. "Formal specification and verification of real-time multi-agent systems using timed-arc petri nets." Advances in Electrical and Computer Engineering 15, no. 3 (2015): 73-8.

[46] Qasim, Awais, Z. U. H. Aziz, Syed Asad Raza Kazmi, Adnan Khalid, Ilyas Fakhir, and Jawad Hassan. "Intelligent agent for formal modelling of temporal multi-agent systems." International Journal on Smart Sensing and Intelligent Systems 13, no. 1 (2020): 1-13.

Effect of Header-based Features on Accuracy of Classifiers for Spam Email Classification

Dr. Priti Kulkarni[1], Prof & Dr. Jatinderkumar R. Saini[2]
Symbiosis Institute of Computer Studies and Research
Symbiosis International (Deemed) University, Pune, India

Prof & Dr. Haridas Acharya[3]
Allana Institute of Management Sciences
Pune, India

***Abstract*—Emails are an integral part of communication in today's world. But Spam emails are a hindrance, leading to reduction in efficiency, security threats and wastage of bandwidth. Hence, they need to be filtered at the first filtering station, so that employees are spared the drudgery of handling them. Most of the earlier approaches are mainly focused on building content-based filters using body of an email message. Use of selected header features to filter spam, is a better strategy, which was initiated by few researchers. In this context, our research intends to find out minimum number of features required to classify spam and ham emails. A set of experiments was conducted with three datasets and five Feature Selection techniques namely Chi-square, Correlation, Relief Feature Selection, Information Gain, and Wrapper. Five-classification algorithms-Naïve Bayes, Decision Tree, NBTree, Random Forest and Support Vector Machine were used. In most of the approaches, a trade-off exists between improper filtering and number of features. Hence arriving at an optimum set of features is a challenge. Our results show that in order to achieve the objective of satisfactory filtering, minimum 5 and maximum 14 features are required.**

Keywords—Email classification; Chi-Square; correlation; relief feature selection; wrapper; information gain; Naive Bayes; J48; spam; support vector machine; random forest; NBTree

I. Introduction

Email communication has become an essential part of all spheres of personal life as well as professional life. But all the emails are not relevant for every user. Day by day the email traffic is increasing, making it imperative to filter spam emails. According to a survey conducted by Radicati 2017, total emails sent and received per day would reach to 319.6 billion by the end of year 2021 [1]. As per Infocomm survey 2016 for internet usage, 'Sending and receiving emails' (94%) and 'Information Search' (92%) are two main activities on internet [2].

Spam finds the first mention as early as in 1975in RFC 706 by John Postel. According to RFC 2505, mass unsolicited emails, sent in large volumes to target the consumers, are called spam emails. Text Retrieval Conference (TREC) defines spam as "unsolicited, unwanted email sent indiscriminately, directly or indirectly, by a sender having no current relationship with the user"(spam track) [10],[11]. According to survey conducted by GFI software in 2014, spam emails consume bandwidth and detract the user from the work. The purpose of sending spam differs from a person-to-person and organization-to-organization. It is used to send phishing, advertising emails or to spread viruses and worm.

An email contains headers and body. Email header field format is defined in RFC 822, RFC 2822. One may classify an email inspecting the body content and headers. Email header contains useful information (Metadata). Contents of the body can be a text, pictorial data, or even sound. This is a purely unstructured part of the email. Our work intends to find out:

a) The minimum number of features in header, necessary to identify spam email.

b) The effect of identified features on the accuracy of classification of email.

c) The best combination of features selection technique and classification algorithm.

This paper consists of three sections, in the first section different approaches for spam classification, as found in literature, are discussed. The next section presents the details about data collection and experiments carried out in for this research. The discussion on results of the experiments follows, along with the conclusion.

II. Literature Review

There are four commonly used techniques for spam classification namely,

a) Use of blacklist [14]

b) Protocol-based approach

c) Use of keywords or content filtering

d) Header based [20],[28],[21],[5],[36],[13]

In the first case, a list of email the network administrator maintains addresses or domain name databases. The classifier matches new record with blacklisted database and simply rejects some mails and puts them onto the spam folder. However, this blacklist requires continuous updation of list. The blacklist approach may fail if the sender's address is fake [37]. The Second approach is protocols based where traffic coming from specific IP address can be blocked. But IP addresses can be easily forged [17], [6]. In the third method of keyword or content filtering [16], spammers bypass the filter by embedding text into images. Such models provides better filtering, however it come with two disadvantages,

a) It is time consuming.

b) The process is language dependent [9].

That is why this paper is focused on the fourth approach of header based filtering. Spam classification helps us to filter the unwanted emails from the email Inbox. There have been various attempts to classify the spam email based on using email header [20],[21],[5],[36],[37],[38],[13],[4], using email body [3],[41],[35],[29],[27],[30],[7],[31],[32],[33],[34] and also using both body and header [18],[23],[21],[15],[42] and statistical features [19],[25]. The email header classification is performed using techniques such as Naïve Bayes (NB), Decision Tree (DT) [40][43], and Support Vector Machine (SVM) [23],[24],[20],[13],[26] Random Forest (RF) [4],[13]. When these techniques were adopted by the researchers using various features and datasets, Random Forest showed better performance than the other techniques. Selecting appropriate set of features is important because that influence accuracy of classifier [8]. Author in [5] used total 26 features derived from behaviour from headers and syslog of emails with back-propagation neural networks (BPNN) and achieved accuracy of 99.6%. But one of the drawbacks of using BPNN is its unstable time to Converge. The number of features and training data affects the performance of BPNN. So the results can fluctuate. In [13] authors have used IP address and subject with other four features which resulted into accuracy of 96.7%. But IP address may get forged. So we have not considered IP address in this research. Our attempt is to suggest optimized features without use of any text data from subject and Body of email. Therefore, we have use combination of different features from literature and by study of personal spam data.

III. Research Methodology

The experiments are conducted in two phases; first Feature Selection techniques are applied on datasets which generate subset of features. In second phase, the resultant feature subsets are used for classification to find the effect on accuracy of classifier. The minimum number of features with classifier is selected as result.

The steps are as follows,

1) Input: Email datasets.

2) Extract Email header features.

3) Apply feature selection techniques.

4) Select subset of features generated by feature selection techniques.

5) Apply classification on Email datasets with selected feature subsets.

6) Classify email into spam and non-spam.

7) Note down the accuracy of the classifiers.

IV. Data Collection and Pre-Processing

We collected emails as reference database to carry the necessary experiments. These emails were collected from personal email account during a period of last 7 years. The two Benchmark corpora available publicly, namely Spam Assassin Corpus and CSDMC2010 corpus are also used in this experiments. These datasets contains spam and ham files. Description of data collected for experimental purpose is given in Table I.

TABLE. I. Description of Email Databases used in Experimentation

Sr. No	Data Set	Description	Spam	Ham	Total
1	S1	Personal Emails	1845	4687	6532
2	B1- Spam Assassin	Benchmark Databases	500	250	750
3	B2-CSDMC 2010		1378	2949	4327

A. Use of Features in Spam Classification

RFC 822 and RFC 2822 are the standard formats, which define email structure and various email header fields. Therefore, the email header field as features are adopted from the above two. The list of features was obtained by study of personal database and from literature. Some of the earlier researchers have not addressed the following six header fields:

- Content-Transfer-Encoding,
- Authentication-Results,
- Presence of ?,!symbols in 'from',
- Presence of ? symbols' in Reply-To' and
- Presence of ? and = symbols in 'Subject'
- presence of $symbol in message-id

So in this experiment an attempt is to rectify the situation, by considering above features.

The features are grouped into two categories,

1) Base features: The features, which are used directly from definitions given in RFCs; as specified in the following list.

Let, $S(U\_B_f)$= { To, Bcc, CC, Received, Return-path, From,Subject,Received-SPF,Authentication-Results,Message-ID, Reply-To, X-Mailer, Content-Transfer-Encoding }.

In further discussions the term $S(U\_B_{f\,)}$ (set of universal base features) is used to refer to above ten features, for the ease of explanation.

2) Derived features: The features, which are constructed from, base features

$S(D_f)$ = {BCC_notempty_To_empty, Message-ID_domainname, Received-Count, Reply-To_domain, Return_Path_Domain, Span_time, Total_Recp,}

In further discussions the term $S(D_f)$ (set of derive features) is used to refer to above seven features ,for the ease of explanation,

Set of features used for experiment by combining base features and derived features is:

S(f)=S(U_Bf) US(Df)

{Authentication-Results, BCC_notempty_To_empty, Content-Transfer-Encoding, From, Message-ID, Message-ID_domainname, Received-Count, Received-SPF, Reply-To_domain, Reply-To, Return_Path_Domain, Return_Path, Span_time, Subject_symbol, To, Total_Recp, X-Mailer}

Table II shows the list of features along with its descriptions used in this study.

TABLE. II. THE LIST OF NUMBER OF FEATURES, ALONG WITH THEIR DESCRIPTION, SELECTED FROM LITERATURE AND BASED ON THE STUDY OF OUR DATASET

Base features involved	**Values extracted**	**Derived feature label**	**Description**	**Reference**
To	To is Empty		Check value of "To" header field exists or if it contains "Undisclosed Recipients" or "<>" symbol	[37], [44]
	To is Undisclosed			[36],[9]
	To contains <>			Proposed feature
BCC,TO	To is empty and BCC is not Empty	BCC_not empty_ To_empty	Check if "BCC" contains email address and "To" do not have any email address.	[36]
To	To_number of address	Total_Recp (To+CC+ BCC)	Total number of email addresses in "To" field	[13] [44]
CC	CC_number of address		Total number of email addresses in "CC" field	[45]
BCC	BCC_number of address		Total number of email address in "To" field	[4]
Received	Count number of received fields	Received_ count	Contains total number of "Received" fields	[4]
Received	Time difference between first received field and last received field, extracted time converted into UTC	Span time	Total travelling time of email from source machine to destination machine.	[4]
From	From contains ?		Check for presence of ?,!,<> symbols in from header field.	proposed feature
From	From contains!			
From	From contains <>			
Subject	Subject contains ?	Subject_ symbol	Check subject field contains symbol "?,="	proposed feature
Subject	Subject contains =			
Base features	**Values extracted**	**Derived feature label**	**Description**	**Reference**
Received-SPF	Received-SPF="bad"		Check Received-SPF field for values as bad, softfail, fail, bad,	[12]
Received-SPF	Received-SPF="softfail"			
Received-SPF	Received-SPF="fail"			
Authentication- result	dkim="bad"		Check Authentication field, dkim value which allows to check email came from authentic domain	proposed feature
Authentication- result	dkim="softfail"			
Authentication- result	dkim="fail"			
Message-id, From	domain name	Message-ID_From_domainname	Check domain in "From" and "Message-id" are not same	[4], [44]
Message-id	Dollar symbol present		Check if message id contains any $ symbol	proposed feature
Reply-To	Reply-To is Empty/exists		Check "Reply -To" is exists or contains"?"	[44]
Reply-To	Reply-To is "?"			proposed feature
Reply-To	Reply-To _domain	Reply-To _domain	Check domain in "From" and "Reply-To" are not same	[44]
X-Mailer	X-Mailer_exist		Check whether X-Mailer exists & checks for valid value of X-Mailer	[4], [13]
Content- Transfer-Encoding	Content-Transfer-Encoding is exists		Check if content transfer Encoding exists/contains no value.	proposed feature
Return-Path	return-path=" " or return-path="bounce"		Check values in "Return-Path" if exists, check if it contains "bounce" word	[44]
Return-Path, From	Return path is NOT matching with From address	Return path_From Domain	Check domain in "return_path" and "From" are same	[45]

V. Experiment

As mentioned earlier, experiments were conducted on three dataset emails as described in Table I. A code is developed in python to extract email header data according to Table II. Our proposed model evaluates email using these 17 features. Each feature is assigned score of 1 (one) if condition is satisfied otherwise it is marked as 0 (zero). The sum of scores was calculated in the end. In this experiment, chi-squared [19], correlation based Feature Selection[39], Information Gain, and relief [22] and Wrapper Feature Selection techniques are applied to find significant features of an email. Classifiers namely Naïve Bayes, Decision Tree, Random Forest, NBTree, Support Vector Machine were used in the experiment.

The data mining tool Weka has been used for applying the machine learning techniques. All the Feature Selection methods and classifiers were adopted in Weka as a selectable runtime parameter. Collected data were arranged in a CSV file in the following format: feature 1, feature 2, feature 3, feature n, class label (Class label indicating two classes, Spam and Ham.) 10 fold cross validation technique is used for data validation. This method uses 90% of the data for training and 10% for testing.

The average weight of each feature generated by all Feature Selection techniques is calculated and listed in Fig. 1.

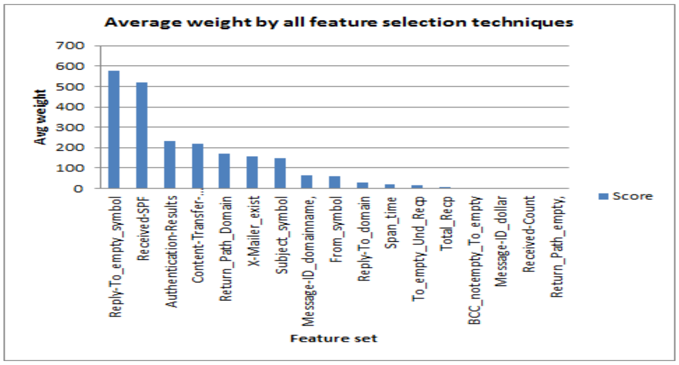


Fig. 1. Average Weight by all Feature Selection Techniques.

It can be clearly observed that our proposed features namely content-transfer-encoding, and Authentication-result, belong to the first five features by weight and have significant contribution to spam classification. The next two features Subject_symbol and From_symbol are among the top ten features. However, our proposed features namely BCC_notempty_To_empty and Message-ID_dollar do not have any significant contribution in spam classification.

VI. Results and Discussion

In this experiment, we have not considered any text feature value from either body or subject. Following are the conventions used in Table III, Table IV and Table V.

FSM1-Chi Squared Feature Selection; FSM2-Correlation based Feature Selection, FSM3- Information Gain, FSM4-Relief Feature Selection, FSM5- Wrapper Feature Selection.

Classifiers:

NB-Naïve Bayes, DT-Decision Tree, RF -Random Forest, NBTree-Naïve Bayes Tree, SVM-Support Vector Machine

As Table III indicates, for dataset S1, the results showed accuracy of 93.53**%** with 17 header features. The maximum features are generated by Relief technique (RT), i.e. 14 features. It maintains best balance between false positive rate and true positive rate. Accuracy of RF is improved by 0.03% with 14 features. With accuracy of 93.56%, Random Forest (RF) outperformed the other four classifiers--Naïve Bayes (NB), Decision Tree (DT), NBtree and Support Vector Machine (SVM).Further, Next to RF, DT classifier also performs well. Naïve Bayes shows stable performance when features are increased from 11 to 14. As number of features reduced, performance of DT and RF decreased. When number of features varied between 11 and 14, Support Vector Machine performed well. However when features are reduced from 11 features to five features, performance of Support Vector Machine decreased by 0.9%.

TABLE. III. Performance of Feature Selection Techniques on the Accuracy of Classifiers on Dataset S1

FSM	**No of features selected**	**NB**	**DT**	**RF**	**NBTree**	**SVM**
**	17	89.72	93.24	**93.53**	91.68	90.53
FSM1	11	89.72	92.39	92.71	91.49	90.54
FSM2	5	89.86	90.45	**90.65**	90.65	90.45
FSM3	11	89.72	92.39	92.71	91.49	90.54
FSM4	14	89.72	93.25	**93.56**	91.47	90.54
FSM5	13	89.72	93.25	93.53	91.78	90.54

TABLE. IV. Performance of Feature Selection Techniques on Accuracy of Classifier on Database B1

FSM	**No of features selected**	**NB**	**DT**	**RF**	**NBTree**	**SVM**
	17	79.33	85.2	85.73	80.8	79.46
FSM1	6	79.46	81.86	81.33	80.93	76.13
FSM2	5	80.66	81.06	**81.33**	80.66	76.13
FSM3	6	79.46	81.86	81.33	80.93	76.13
FSM4	12	80.66	83.6	**83.73**	80.93	77.2
FSM5	7	80.26	81.06	80.66	80.66	77.06

TABLE. V. Performance of Feature Selection Techniques on Accuracy of Classifier on Database B2

FSM	**No of features selected**	**NB**	**DT**	**RF**	**NBTree**	**SVM**
**	17	72.48	93.28	94.71	93.42	85.73
FSM1	14	79.57	90.67	91.06	91.01	85.41
FSM2	5	70.58	85.66	**86.17**	86.02	84.69
FSM3	14	79.57	90.67	91.06	91.01	85.41
FSM4	13	72.48	93.28	**94.78**	93.81	85.73
FSM5	8	71.53	93.44	71.53	93.51	84.71

Moreover, Correlation based Feature Selection technique generated five features, which are minimum number of features. When features are reduced from 17 to 5, accuracy of Random Forest (RF) is reduced by 2.36%. In short, RF and NBTree classifiers give high accuracy of 90.65% as compared to the other three. With five features, accuracy of NB is the lowest among all. However, even otherwise, NB did not perform so well as other four classifiers even with more number of features.

On benchmark dataset B1 of the size of 750 data records, Random Forest performs best (83.73% accuracy) with maximum of 12 features. Correlation based Feature Selection method generated 5 features, the minimum number in this dataset resulting into accuracy of 81.33% with RF classifier. On benchmark dataset B2, RF shows better performance, giving accuracy of 94.78% as compared to other two datasets. In this dataset also, relief Feature Selection technique with RF classifier outperformed others even with 13 features. In the same way, Random Forest performed better with accuracy of 91.6% when Chi Square and IG generated 14 maximum features. Correlation based FS generates 5 features. With minimum number of 5 features accuracy reduce by 4.89%. One of the common observations is that Random Forest method with Relief as Feature Selection technique performs better on all the datasets.

Following are the set of minimum and maximum number of features:

S_min_5={Total_Recp, Subject, Received-SPF, Authentication-Results, Reply-To}

S_max_14={Authentication Results, BCC_notempty_To_empty, Content-Transfer-Encoding, From_symbol, Message-ID_domainname, Received SPF, Reply-To_domain, Reply-To_empty_symbol, Return path_FromDomain, Span_time, Subject_symbol, X-Mailer_empty, To_empty_Und_Recp, Total_Recp}

VII. Conclusion

In this paper, we evaluated performance of five Feature Selection techniques and five classifiers on email headers. Our header based approach for Feature Selection showed that minimum five features generated by correlation based Feature Selection technique performed well on all three datasets with varying accuracy 70.58% to 90.65%. Relief Feature Selection technique generated the maximum fourteen features with varying accuracy of 91.06% to 94.78%.This implies that the features we proposed namely, Authentication-result and content-transfer encoding play significant role in identifying spam emails. The result of our experiment result shows that Random Forest performs better than all other classifiers in terms of accuracy as well as number of features.

References

[1] The Radicati Group, Inc. https://www.radicati.com/wp/wp-content/uploads/2017/01/ Email-Statistics-Report-2017-2021-Executive-Summary.pdf

[2] Annual Survey On Infocomm Media Manpower For 2016(https://www.imda.gov.sg/-/media/imda/files/industry-development/fact-and-figures/infocomm-survey-reports/infocomm-media-manpower-survey-2016_public-report.pdf?la=en)

[3] Ayodele T., Zhou S., Khusainov R.: Email Classification Using Back Propagation Technique, International Journal of Intelligent Computing Research, 2010.

[4] Al-Jarrah, O., Khater, I., & Al-Duwairi, B. (2012). Identifying potentially useful email header features for email spam filtering. In The Sixth International Conference on Digital Society (ICDS) (Vol. 30, p. 140).

[5] C.-H. Wu, “Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks,” Expert Systems with Applications, vol. 36, pp. 4321-4330, April, 2009

[6] Chirita, P. A., Diederich, J., & Nejdl, W. (2005, October). MailRank: using ranking for spam detection. In Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 373-380). ACM.

[7] Daeef, A. Y., Ahmad, R. B., Yacob, Y., Yaakob, N., & Azir, K. N. F. K. (2016). Multi Stage Phishing Email Classification. Journal of Theoretical & Applied Information Technology, 83(2).

[8] De Stefano C, Fontanella F, Marrocco C, et al. A GA-based Feature Selection approach with an application to handwritten character recognition. Pattern Recognition Letters 2013

[9] F. Salcedo-Campos, J. Dнaz-Verdejo, P. GarcнaTeodoro, Segmental parameterization and statistical modelling of e-mail headers for spam detection. Information Sciences, no. 195(2012), 2012, pp. 45-61.

[10] Gordon Cormack and Thomas Lynam.(2005). Spam corpus creation for TREC. In Proceedings of Second Conference on Email and Anti-Spam, CEAS’2005,

[11] Gordon Cormack, TREC 2006 Spam Track Overview (http://trec.nist.gov/pubs/trec15/papers/SPAM06.OVERVIEW.pdf)

[12] Gorling, S. (2007). An overview of the Sender Policy Framework (SPF) as an anti-phishing mechanism. Internet Research, 17(2), 169-179.

[13] Hu, Y., Guo, C., Ngai, E. W. T., Liu, M., & Chen, S. (2010). A scalable intelligent non-content-based spam-filtering framework. Expert systems with applications, 37(12), 8557-8565.

[14] Jung J, Sit E. An empirical study of spam traffic and the use of DNS black lists. In: Proceedings of fourth ACM SIGCOMM conference on internet measurement, Taormina, Sicily, Italy; October 2004.

[15] Jason Chan, Irena Koprinska, Josiah Poon, Co-training with a Single Natural Feature Set Applied to Email Classification", In:Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence 2004

[16] Kelly Jackson Higgins, Dark Reading. Botnets Battle Over Turf. http://www.darkreading.com/document.asp? doc id=122116, Apr. 2007

[17] Ramachandran, A., Feamster, N., & Vempala, S. (2007, October). Filtering spam with behavioral blacklisting. In Proceedings of the 14th ACM conference on Computer and communications security (pp. 342-351). ACM.

[18] Kiritchenko S., Matwin S., Abu-Hakima S. (2004) Email Classification with Temporal Features. In: Klopotek M.A., Wierzchon S.T., Trojanowski K. (eds) Intelligent Information Processing and Web Mining. Advances in Soft Computing, vol 25. Springer, Berlin, Heidelberg

[19] Gomez, J. C., Boiy, E., & Moens, M. F. (2012). Highly discriminative statistical features for email classification. Knowledge and information systems, 31(1), 23-53.

[20] M. Ye, et al., “Spam Discrimination Based on Mail Header Feature and SVM,” In Proc. Wireless Communications, Networking and Mobile Computing, 2008. WiCOM ’08. 4th International Conference on Dalian Oct. 2008

[21] Jyh-Jian Sheu, Ko-Tsung Chu, Nien-Feng Li, Cheng-Chi Lee, 2017, An efficient incremental learning mechanism for tracking concept drift in spam filtering, PLoS one

[22] Kira, Kenji and Rendell, Larry (1992). The Feature Selection Problem: Traditional Methods and a New Algorithm. AAAI-92 Proceedings.

[23] Lai, C. (2007). An empirical study of three machine learning methods for spam filtering. Knowledge-Based Systems, 20(3), 249–254.

[24] Wang, M. F., Jheng, S. L., Tsai, M. F., & Tang, C. H. (2011, July). Enterprise email classification based on social network features. In 2011

International Conference on Advances in Social Networks Analysis and Mining (pp. 532-536). IEEE.

[25] Martin, S., Nelson, B., Sewani, A., Chen, K., & Joseph, A. D. (2005, July). Analyzing Behavioral Features for Email Classification. In CEAS

[26] Patidar, V., Singh, D., & Singh, A. (2013). A Novel Technique of Email Classification for Spam Detection. International Journal of Applied Information Systems, 5(10).

[27] Ruan, G. and Tan, Y. (2010). A Three-Layer BackPropagation Neural Network for Spam Detection using Artificial Immune Concentration. Soft Computing, 14(2), pp. 139-150.

[28] Sheu J.J, “An Efficient Two-phase Spam Filtering Method Based on E-mails Categorization “ International Journal of Network Security, vol.9, pp. 34-43, July 2009.

[29] Senthamarai Kannan Subramanian and N. Ramaraj, 2007. Automated Classification of Customer Emails via Association Rule Mining. Information Technology Journal, 6: 567-572.

[30] Schmida M. R., Iqbalb F., Fungc B. (2015) E-mail authorship attribution using customized associative classification. The Proceedings of the Fifteenth Annual DFRWS Conference, Volume 14, Supplement 1, August 2015, Pages S116–S126.

[31] Sahın, E., Aydos, M., & Orhan, F. (2018, May). Spam/ham e-mail classification using machine learning methods based on bag of words technique. In 2018 26th Signal Processing and Communications Applications Conference (SIU). IEEE.

[32] Saini, J. R., & Desai, A. A. (2010). “Analysis of Classifications of Unsolicited Bulk Emails”. International Journal of Computer and Information Engineering, 4(1), 115-119.

[33] Saini J. R., Desai A. A.(2010), “A Survey of Classifications of Unsolicited Bulk Emails”, in National Journal of Computer Science and Technology (NJCST), 2010. ISSN 0975-2463

[34] Saini J. R.,Desai A.A,(2010), “A Supervised Machine Learning Approach with Re-training for Un-structured Document Classification in UBE”, published in INFOCOMP Journal of Computer Science; ISSN: 1807-4545.

[35] Trevino, A. 2007. Spam Filtering Through Header Relay Detection.

[36] Wang C-C. Sender and receiver addresses as cues for anti-spam filtering. Journal of Research and Practice in Information Technology 2004;36(1):3–7.

[37] Wang, C. C., & Chen, S. Y. (2007). Using header session messages to anti-spamming. Computers & Security, 26(5), 381-390.

[38] W. Li, W. Meng, Z. Tan, and Y. Xiang,(2014) "Towards designing an email classification system using multi-view based semi-supervised learning," in 13th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2014, 2015, pp. 174-181.

[39] Yang, Q., & Gras, R, 2010, December. How dependencies affect the capability of several Feature Selection approaches to extract the key features. In Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on(pp. 127-134). IEEE.

[40] Youn, S., & McLeod, D. (2007). A comparative study for email classification. In Advances and innovations in systems, computing sciences and software engineering (pp. 387-391). Springer, Dordrecht.

[41] Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998, July). A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization: Papers from the 1998 workshop (Vol. 62, pp. 98-105).

[42] Zhang, L., Zhu, J., & Yao, T. (2004). An evaluation of statistical spam filtering techniques. ACM Transactions on Asian Language Information Processing (TALIP), 3(4), 243-269.

[43] Ying, K. C., Lin, S. W., Lee, Z. J., & Lin, Y. T. (2010). An ensemble approach applied to classify spam e-mails. Expert Systems with Applications, 37(3), 2197-2201.

[44] Qaroush, A., Khater, I. M., & Washaha, M. (2012). Identifying spam e-mail based-on statistical header features and sender behavior. In Proceedings of the CUBE International Information Technology Conference (pp. 771-778). ACM.

[45] Zhang, D. Y., & Yang, L. (2014). Implementation of Mail Classification Using Neural Networks of the Second Type Spline Weight Functions. In Applied Mechanics and Materials (Vol. 513, pp. 687-690). Trans Tech Publications Ltd.

Enhancing the Quality of Service of Cloud Computing in Big Data using Virtual Private Network and Firewall in Dense Mode

Hussain Shah[1]
Department of Computer Science
Islamia College University,Peshawar, KP, Pakistan
Institute of Computer Science and Information Technology
University of Agriculture, Peshawar, KP, Pakistan

Aziz ud Din[2]
Shaykh Zayed Islamic Centre
University of Peshawar, Peshawar, KP, Pakistan

Abizar[3], Shams ud Din[5]
Department of Computer Science
Islamia College University, Peshawar, KP, Pakistan

Adil Khan[4]
Department of Computer Science
University of Peshawar, Peshawar, KP, Pakistan
Shaykh Zayed Islamic Centre
University of Peshawar, Peshawar, KP, Pakistan

***Abstract*—Cloud Computing entails accessing and storing programs and data over the internet instead of the hard drive of a personal computer. Over the Internet, it is the practice of software and hardware to pass a service. Cloud gives the ability to consumers to access big data and use applications from every device that can have access to the internet, however, the key problem is security and this can be solvable by a firewall and Virtual Private Network. Recently, research has been accomplished in deploying firewalls and Virtual Private Networks with parameters of throughput and load in sparse mode. In this paper, an examination of firewall and Virtual Private Network is considered based on average throughput, average packet loss and average end-to-end delay in dense mode. To examine the performance of cloud computing without Firewall and Virtual Private Network, with firewall only, and with firewall and Virtual Private Network is the research goal. The simulation results have shown that Firewall and Virtual Private Network offers better security through a wide investigation with slight distress in the cloud performance.**

Keywords—Cloud computing; big data; firewall; virtual private network; security; performance

I. Introduction

The Internet is growing vigorously these days. The cost of storing data, the power consumed by computers, and the hardware are increasing and expanding. The storage space in the data centre isn't enough to meet the requirements. Also, the system and service of the internet can't solve the said issues. The researchers and academia work to find new solutions. At the same time, large enterprises have to study data sources entirely to support their business. The collection and analysis must be built on a new platform such as Cloud Computing. In [1], the need for Cloud Computing by the business community is addressed? It is stated how to utilize the resources of a computer, how to increase the economic efficiency by improving the utilization rate, and how to decrease the equipment energy consumption. Cloud Computing is a computing technique in which capable and changeable information technology (IT) gives service to external clients using internet technology. Cloud Computing is not a fundamental idea instead it's a developmental concept that combines different existing techniques to recommend a new useful IT providing tool. Through the internet, Cloud applications expand their availability and accessibility by using large data centres and powerful servers that host web applications and services [2]. Those who have a standard internet connection, as well as browser, can be connected to the cloud applications. Cloud-based computing is a model that allows suitable on-demand network access to a shared pool of configurable assets of computing resources (e.g., networks, servers, storage, services, applications) that could be quickly provisioned and released with minimal management efforts or service provider interaction [3].

Cloud Computing technique helps in computing different tasks like efficiency, reduced cost, performance, quick deployment and easy access to the information, etc. The important issue in cloud computing is the security which needs to be improved. Earlier a couple of security mechanisms such as firewall and VPN has already been introduced and standardized for guaranteeing the security that influences the cloud performance in regards to the quality of service parameters. As per the literature survey, very few research attempts have been observed and prepared to examine average end-to-end delay, average throughput and average packet loss in dense mode.

In this paper, the research is carried out on VPN and firewall in a dense mode that base on the average throughput, average end-to-end delay and average packet loss using the OPNET 14.5 simulator.

A. Applications of Cloud Computing

Cloud Computing provides several benefits to cloud users where one of the application is presented below in Fig. 1 for more observation.

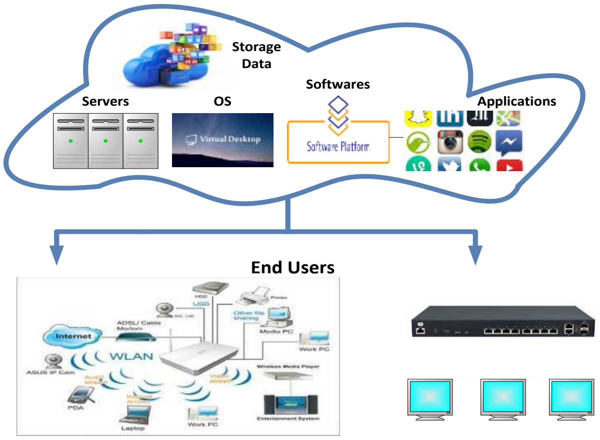


Fig. 1. A Cloud Application.

1) E-Learning: Cloud Computing is a significant technique that can be used in education (e-learning) to create attractive environments for teachers, students, and researchers to retrieve information by using the cloud of parent organization [4].

2) E-Governance: A government can provide an efficient service to its citizens, institutions, and their cooperation by using Cloud-based computing [5]. This can make the environments more scalable and customizable by reducing the energy to manage, install, and upgrade the applications.

3) Cost efficiency: By using the Cloud Computing technique, the Cost and budget of a company can be reduced to a great extent rather than relying purely on traditional desktop-software based approaches [6], such as it provides the facilities to users or customers to use hardware and software owned by other companies without the hesitant of managing and purchasing them, or without purchasing the required application by accessing the third party servers with the help of internet.

4) Almost unlimited storage: Unlike traditional desktop-based computing approaches. Cloud computing offers the facility of unlimited storage.

5) Backup and recovery: As compared to physical desktop hard drives, in cloud computing the data is stored on many servers across the globe where one can easily retrieve data from the cloud [7].

6) Easy access to information: By accessing the cloud one can easily upload and download data from anywhere in the world by using different gadgets.

7) Automatic software integration: Cloud-based computing is the automatic integration system that can integrate and update the software automatically which means that there is no need of the user to update the software itself.

8) Quick deployment: One of the vital advantages of cloud computing is its fast deployment. Once the account and procedure of data uploading and downloading are familiarized, then the user can easily retrieve data anywhere by using the application with the appropriate support of internet connection.

9) Fresh software: With the help of SaaS (Software as a Services), Cloud computing provides the latest version of software's to use in commerce and also to clients when they are released [8].

10)Always-on availability: The cloud providers are trustworthy to deliver their services and facilities to users/customers as efficiently as possible.

B. The Architecture of Cloud Computing

From the architecture perspective, the cloud computing architecture is composed of several important characteristics or components, three service models and five deployment models, that are illustrated in Fig. 2.

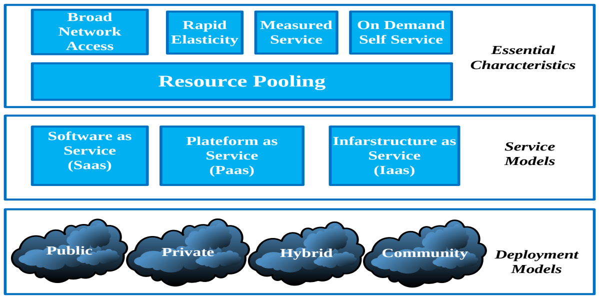


Fig. 2. The Architecture of Cloud Computing.

1) Deployment models of cloud: Cloud-based computing is established [9] on settlement models such as Public Cloud, Private Cloud and Hybrid Cloud, moreover, for different purposes, community cloud networks and mobile cloud are also used.

a) Public Cloud: This model delivers and stores a huge size of data and other facilities for the access of the general community from facility providers, spending facilities as pay/ use or cost-free.

b) Private Cloud: This model is used in fog computing by a different organization and recycled via the certified worker of that organization. In other words, Private cloud is one of the deployment models that is normally used by an individual organization or used by the authorized users of that organization.

c) Hybrid Cloud: In such systems, the organization use the important data or information on the private cloud, and the data which is less secured is being used on public could thus in such a situation the Hybrid Cloud is commonly preferred to be used, which means that this model is the mixture of double models of Cloud deployment such as community, private or public cloud models.

d) Community Cloud: A community cloud is a distinctive model of cloud deployment in which an organization is dispersed by numerous organizations that chain a precise community that has mutual concerns. A community cloud is shaped when numerous organizations share common infrastructure with similar necessities.

e) Mobile Cloud: The practice of cloud computing in mixture with portable mobile devices is known as being a Mobile cloud [10]. The occurrence of Cloud computing happens when on the internet data and information are kept somewhat compared to separate strategies, giving access on demand. In the situation of mobile cloud, applications run on the server remotely and formerly user receives them. Mobile applications are rapidly developing a section of the worldwide mobile market. Several mobile corporations have their cloud and the user takes functionality from the mobile cloud.

2) Services model of cloud: Cloud computing service suppliers provide the three services to the end-user such as Infrastructure as a Service (IaaS), Software as a Service (SaaS), Platform as a Service (PaaS) [11].

a) Infrastructure as a Service (IaaS): Infrastructure as a Services (IaaS) provides physical components over the internet. Infrastructure as a Services provides the Infrastructure physically or virtually such as load balancer, virtual machine, data storage spaces, caching. IaaS is known as Hardware as a Service, as in IaaS the clients aren't afraid of managing and purchasing data centres and hardware's as all these things are controlled by the cloud service provider. The Cloud services provider could allow one to store data in the data center based on the size requested. A few examples of IaaS include Google drive, VMware and Rack space.

b) Software as a Service (SaaS): It permits customers to execute the available online application and software. These are retrieved over the internet. Such types of the platform in cloud computing transfer programs to millions of clients or user over the browser. SaaS is normally employed in the human resource management system and ERP (Enterprise Resource Planning). Google Applications and Zoho Office are giving such services too. Microsoft Word, Google docs, Facebook, Twitter, Gmail, and Google Calendar are some other examples of SaaS.

c) Platform as a service (PaaS): Platform as a service lets (PaaS) is a variety of cloud computing that provides a platform for the user to build and install some fresh applications online. Such as to build software or website. The main goal of PaaS is to develop, deploy and test the code effortlessly. A few examples of PaaS are Engine Yard, Force.com, Google App Engine, Apache Stratos, Azure, and Yahoo Pipes.

C. Security Issues in Cloud Computing

Cloud computing comes up with numerous major issues and trials concurrently. Like availability, performance, and security. In [12], amongst the challenges in Cloud computing, security is one of the significant and critical issues.

The security challenges in Cloud-based computing are very vast, dynamic and versatile [13]. Location transparency and data location is an important issue in the security of Cloud computing as the record is stored on virtual servers in the cloud. The users without knowing the exact location of its data storage due to which the act about data protection might be violated and affected.

In Cloud-based computing, security issues occur due to the usage of the network in Cloud computing, as users want a network connection to enter that information and resource that is of need [14]. Due to which an unauthorized user may also interfere in the network of Cloud computing. As shown in Fig. 3, the security issues are rated up to 74.4% amongst all the challenges faced by Cloud Computing.

The main issue in Cloud-based computing is to assure security. Therefore, a security technique needs to be deployed that permits only those users who are authorized and blocks those users who are not trustworthy in the cloud computing network. Two methods or techniques are deployed in an association such as firewall and VPN to improve the security in Cloud-based computing. VPN is one of the preferred technique that is used for secure data transmission from and to the Cloud. Within the VPN secured and reserved sub tunnels can be generated. VPN connects and transmits data with the help of a concept called tunneling. First, the packet is protected (encapsulated) in a fresh packet by a fresh header before it is transmitted into the VPN tunnel. The header provides information about the router of the corresponding packet, while the packet is roaming in a network that is shared before it is gotten by the tunnel destination. This encoded track is enveloped or compressed in which the packet travels is known as a tunnel. This summarized packet is 'de-capsulated' and sent to the final destination when it extends the endpoint of the tunnel. Both the termination points of the tunnel desires to provide a similar tunneling protocol. That protocol works on the data link layer (layer two) or the network layer (layer three) of the open system interconnection (OSI) model. The best well-known protocol used for VPN is Internet protocol security (IPsec) and point to point tunneling protocol (PPTP). VPNs are usually employed by using the IPsec. It is a standard way for the employment of a VPN. The IPsec and VPN are recognized very well and developed in a manner to offer strong security which gives access control, data confidentially and authentication. By assimilating IP security infrastructure into the wireless LAN's infrastructure is a simple effort to transmit wireless traffic and the VPN will provide the security to that traffic [15] as shown in Fig. 4.

A firewall is used for packet filtering between the outside world and the internal network. As the firewalls have been employed on large public networks from many years that's why firewalls have been used with VPN. Another reason for using a firewall with VPN is because of its important role and the security of the network. The joint implementation of firewall and VPN has a great impact on the performance of Cloud Computing in terms of quality of service (QoS) parameters [16].

D. Firewall

A firewall is a device used for security that detects incoming and outgoing traffic for network and a choice is made on the basis that which packet needs to be allowed and which to be blocked based on administration policy for the firewall [17]. It is like a barrier and the entire traffic (leaving or arriving) must be passed via this barrier. Only permitted traffic as defined by the cloud service provider in local security policy that will be allowed to pass. The firewall is normally considered as a tool that filters the packets, that acts as a barrier between the public and private networks. The word firewall is used in a computer that implies a device that guards the network against traffic that is untrusted.

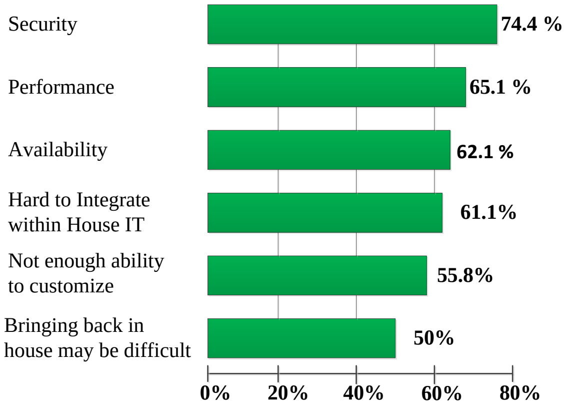


Fig. 3. Challenges to the Cloud Computing.

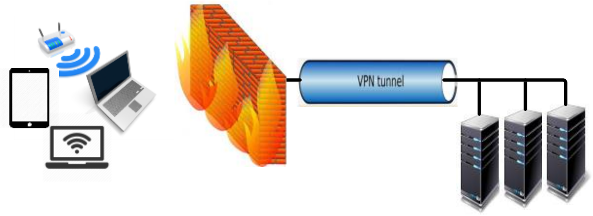


Fig. 4. VPN Procedure within IPSec.

1) Types of firewall: Firewalls are classified into three basic types: proxy servers (that is divided into two subtypes application gateways, circuit-level gateways firewall), state-full packet filters and packet filters firewall [18], as shown in Fig. 5.

a) Packet Filters Firewall: One of the most basic types of firewall is the packet filter firewall. Packet filter is applied for safety to shield the inside network users from outside network threats. This kind of firewall is the initial firewall that is used for the security of the network. It is used to monitor network entrances or access by observing incoming and outgoing packets and then making a decision based on the interior protocol address (IP address) of source and destination to allow and halt packets from the network. This packet filter firewall works on the third layer of the OSI model which deliver highly effective security mechanism. This kind of firewall is also known as static filtering. When it is implemented in a network the packet filtering is one of the most important procedures that are essential for security concern.

b) Proxy Servers: A proxy server is a kind of firewall that saves and shelters the properties of the network by data filtering at the seventh layer of the OSI model. This kind of firewall is the best kind of firewall. It provides security that is improved due to proxy data and information that doesn't allow transmitting over proxy as proxy acts as an intermediate between server and clients. A proxy server firewall provides and delivers internet access to network users. They are either on the application layer or the transport layer. This type of firewall is of two categories one is application gateway (work on application layer) and second is the circuit-level gateway (work on transport layer).

c) State-full Packet Filters: State-full packet filters are similar to a screen that exists between the server and users. This device uses state-full packet filtering for observing all packets of data when arrived on the screen. The screen examines the data based on the set of security policies.

E. Virtual Private Network (VPN)

VPN (Virtual Private Network) technology provides a way of protecting information being transmitted over the internet, by allowing users to establish a virtual private "tunnel" to securely enter in internal networks, accessing resources, data and communications via an insecure network such as the internet. A VPN is a private network connection [19] that provides one's a facility of secure connection in existing public network in a remote area. In VPN each record (video, voice, and file) is an encrypted form goes to a secure virtual tunnel among the clients and the VPN provider server to cloud computing services.

1) VPN tunneling: A VPN tunnel is an encoded or encrypted or cipher path between a user and another network. To learn more that how a VPN works then it easily understand to looking at the procedure of tunneling data. A VPN tunnel is often called a virtual private network which is an encrypted path between one's computer and the server of VPN that provides the VPN services. As the connection is encrypted nobody is allowed to monitor, modify or stop one's communication. All of the communication and the data is travel in a VPN tunnel so nobody is allowing to examine the data. The VPN tunnel protects one's chatting, browsing and all other traffic from the snooping eyes of one's Internet service provider (ISP), government and also from the person who controls the Wi-Fi (wireless fidelity) which one uses to connect as shown in Fig. 6 [20].

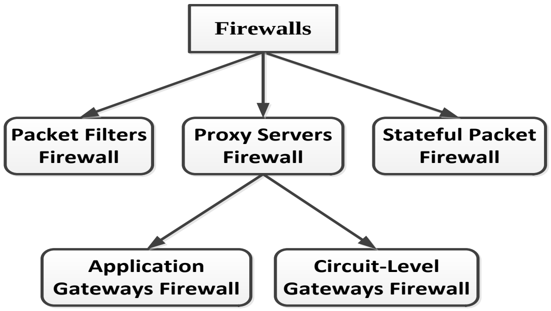


Fig. 5. Different Types of Firewalls.

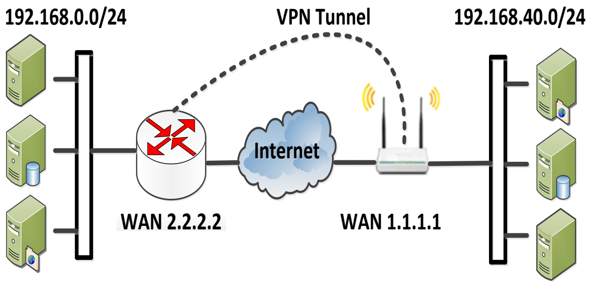


Fig. 6. VPN Tunneling.

2) Privacy in VPN tunnel: A VPN tunnel offers safe and free of intrudes connection [21]. Moreover, using VPN hides the IP address, and browsing data. Nobody can discover your real locality or IP address if one is using a VPN tunnel. One's VPN server will be merely catchable.

II. Related Work

After studying the literature in Cloud computing, different techniques are used for ensuring the security of networks, such as firewall and VPN. Firewall plays a vital role in network security because a firewall can scan all the traffic on the network and filter the packets and allow only those packets or users which are authorized. While implementing a firewall, the network administrator faces issues of conflicting policies. A firewall supports multiple distributed policy which may cause delay, system overhead and time-consuming. Various authors have used firewall and VPN for security purpose however every one has its limitations. In this paper, the issue of security is the main problem in the wireless LAN standard IEEE 802.11 in Cloud computing [22].

The performance is a major issue as the firewall tool checks all incoming and outgoing packets, it consumes time and produces overhead in the system which affects the service level agreement (SLA) [23]. A cloud-based firewall is difficult to configure efficiently. To support distributed processing environments and to overcome the conflict of making security policy rules set by the network administration [24]. Implementing Firewalls cost enough budget as for as low-level business is concerned. The amount of implementing a firewall is approximately 116,075$ for one year to keep its deployment and maintenance [25]. VPN is a security mechanism that allows user to access common applications such as HTTP, load, Email. However, using the VPN can achieve the security but it also degrades the performance of the network in terms of throughput, etc. [26]. Once some attacks occur against Cloud service the response time of system firewall becomes overhead of performance due to the huge arrival of packets. So, it will take a long response time which will be the violation of service level agreement (SLA) and the decrease in customer fulfilment [27]. In a computer networking environment, a firewall protects internal nodes from the external attack and the internal nodes as well because a firewall is managed by the system administrator. Therefore, it is needed to handle the firewall in a new way which satisfies the requirement of Cloud computing [28]. Firewalls can be an essential part to secure network that prevents hackers away from a computer network, in this regard, the procedure of configuring a firewall is a difficult and stressful job [29]. When the external users try to enter the Cloud computing network, so, first they undergo through the vital barrier of firewall that provides networks security and allow only those users who are compliant and give safety from different attacks such as HTTP DoS (Denial of service) or brute force attack.[30, 31].

III. Methodology

A. Simulation Parameter Selections

The parameters used in the simulation are given below in Table I.

TABLE. I. Parameter for Simulation

Parameters	***Description***
Simulation tool	OPNET modular 14.5
Standards	IEEE 802.11g (WLAN)
Time	5 Minutes
Area	100 m x 100 m
Nodes	18
Workstations	03
Access points	03
Servers	03
Protocols	OSPF
Applications	HTTP, FTP, Email

B. Performance Parameters

In this research there are three performance parameters used are discussed as follows.

1) Average throughput: The number of successfully received packets from source to destination as per unit time. It is calculated in bits per second (bps) or packet per second and can be calculated as shown in equation (1).

$$\text{Average Throughput}=\sum_{i=1}^{n}\frac{\text{(Received Packet i x Packet Size)}}{\text{Simulation Time}} \quad (1)$$

2) Average end-to-end delay: It is the total time taken by a packet to reach from source to destination and is represented in seconds/millisecond. Thus, in the research work achieved, one of the parameters is the average end-to-end delay as declared in equation (2).

$$\text{Average End to End Delay} =\sum_{i=1}^{n}\frac{\text{(Packet Received time i - Packet Sent time i)}}{\text{Total number of packets }(n)} \quad (2)$$

3) Average packet loss: Packet loss happens whenever a packet flops to the extent of the target while roaming through a network of computers. It is normally initiated by crowding over a network. In the presence of Firewall and VPN, it is also significant to investigate the average packet loss and can be calculated as stated in equation (3).

$$\text{Average Packet Loss} =\sum_{i=1}^{n}\frac{(\text{Packet Loss}_{\text{i}}\text{ x Packet Size})}{\text{Simulation Time}} \quad (3)$$

C. Network Objects

The following objects are used.

1) Applications Configuration
2) Profile Configuration
3) IP VPN Configuration
4) Wlan wkstn **(**clients**)**
5) ip32 cloud(Internet)
6) ethernet4 slip8_gtwy (router)
7) wireless Ethernet slip4 Router (Access Point) which configured on BSS.
8) Ethernet slip8 firewall (firewall)
9) PPP DS1
*10)*ppp server (server)

D. Network Simulation Scenario

In this paper, the optimize network simulator was chosen that contain three different scenarios that will investigate the performance of the network with different illustration as mention below.

1) Without firewall and VPN scenario: In the scenario shown in Fig. 7, there are several workstations connected to three Access Points (Access Point-1, Access Point-2, Access Point-3) which are configured for three BSS. The Access Points are connected by PPP-DS1 to IP cloud (Internet) and then further connected by PPP-DS1 to Router D connected by PPP-DS1 to three Servers (Server AA, Server BB, Server CC) which represents three departments. The scenario architecture and layout are as shown in Fig. 7.

2) With firewall no VPN scenario: In the scenario shown in Fig. 8, there are several workstations connected to three access points (Access Point 1, Access Point 2, Access Point 3) which are configured for three BSS. The access points are connected by PPP-DS1 to IP cloud (Internet), then further connected by PPP-DS1 to Firewall and Router D by PPP-DS1 to three Servers (Server AA, Server BB, Server CC) which signifies three departments. In the scenario, the firewall is selected to stop servers from any exterior entree to/ (browsing over the web) from the servers.

3) With firewall and VPN scenario: In the scenario shown in Fig. 9, several workstations are connected to three access points (Access Point 1, Access Point 2, Access Point 3) which are configured for three BSS. These access points connected by PPP-DS1 to IP cloud (Internet), then further connected by PPP-DS1 to Firewall and Router D by PPP-DS1 to three Servers (Server AA, Server BB, Server CC) which represents three departments. In the last scenario, the firewall is used to stop servers from any outside access to HTTP (web browsing). The VPN tunnel would be chosen to let the clients (PCs) from Access Point-1 to access HTTP (web browsing) from the servers in the scenario. The traffic generated by Access Point-1 is not cleaned by the firewall and let users from the Access point-1 because the IP packets in the tunnel will be condensed inside an IP datagram. The scenario design and arrangement is as shown in Fig. 9.

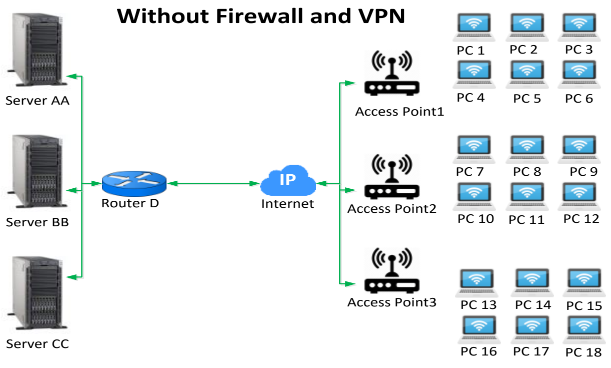


Fig. 7. The Architecture and Layout of Scenario-1.

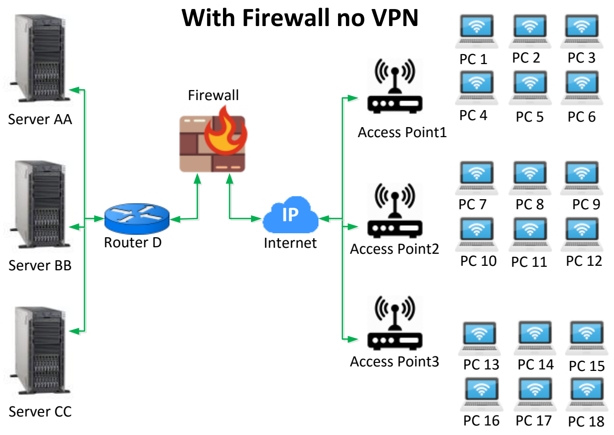


Fig. 8. The Architecture and Layout of Scenario 2.

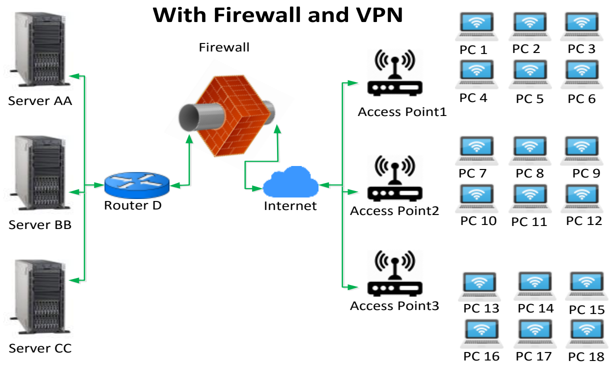


Fig. 9. The Architecture and Layout of Scenario-3.

IV. Comparison and Analysis

After implementing the scenario, the results are packed, stored and compared with each other, then the results are graphed by using the Origen Lab 2020, the parameters selected for the decision are average throughput, average end-to-end delay, and average packet loss.

Three scenarios had been prepared to examine the impact of firewall and VPN in Cloud-based computing in current research by using optimized network modular 14.5 simulators. In the paper, the results and graphs are discussed below for investigating the performance of the Cloud computing network after applying a firewall and VPN.

A. Simulation Results

For each scenario, to inspect the performance of Cloud computing "without Firewall and VPN", "with Firewall no VPN" and "with Firewall and VPN" are used, the following three performance parameters that are 'Average Throughput', 'Average End-to-end Delay', and 'Average Packet loss'.

1) Average throughput: The number of successfully received packets from source to destination as per unit time. It is calculated in bits per second (bps) or packet per second.

Table II expresses simulation results of the average throughput with no Firewall no VPN, with Firewall no VPN and with Firewall VPN in the Cloud computing network.

TABLE. II. RESULT OF AVERAGE THROUGHPUT FOR NO FIREWALL NO VPN, WITH FIREWALL NO VPN AND WITH FIREWALL AND VPN

Time	**Average Throughput (bits/second)**		
	No Firewall No VPN	**Firewall No VPN**	**Firewall VPN**
0	0	0	0
50	2204.44	853.33	852.30
100	1558.58	873.41	843.29
150	80887.11	66030.11	52521.20
200	6547912	50596.71	41936.02
250	56489.87	47249.27	43493.59
300	51708.13	42186.77	40323.41

Fig. 10 shows the comparison of average throughput for “no Firewall and no VPN”, “with Firewall no VPN” and “with Firewall and VPN” in Cloud-based computing. 18 nodes and 3 servers are included in the scenario. The simulation time per second is displayed on the horizontal axis, whereas the network average throughput (bits/sec) is displayed on the vertical axis. The network average throughput presence of no firewall no VPN is represented by the square line while network average throughput presence of with firewall no VPN is showed by circle line, whereas the network average throughput presence of with firewall and VPN is presented by triangle line. The average throughput improved with the presence of nodes ‘without firewall and VPN’ as compared to ‘with Firewall no VPN’ and ‘with the firewall with VPN’ by the wide investigation since without any hurdles users can send and receive the data. The impact of ‘firewall and VPN’ on the cloud computing network is verified and It has been confirmed from the graph that the presence of ‘no firewall no VPN’ gives an improved rate of average throughput than the presence of ‘firewall and no VPN’ and ‘with the firewall with VPN’ in a cloud-based computing network. It was revealed through broad simulation that firewall and VPN affect cloud performance while provides better security.

2) Average end-to-end delay: It is the total time taken by a packet to reach from source to destination and is represented in seconds/millisecond.

Table III expresses the simulation results of average end-to-end delay with no Firewall no VPN, with Firewall no VPN and with Firewall VPN in Cloud computing network.

Fig. 11 shows the comparison of average end-to-end delay for ‘no firewall and no VPN’, ‘with Firewall no VPN’ and ‘with a firewall with VPN’ in Cloud-based computing. 18 nodes and 3 servers are included in the scenario as well. The simulation time per second is displayed on the horizontal axis, whereas the network average end-to-end delay (sec) is displayed on the vertical axis. The network average end-to-end delay (sec) presence of ‘no firewall no VPN’ is represented by the square while network average end-to-end delay (sec) presence of ‘with firewall no VPN’ is showed by circle line, whereas the network average end-to-end delay (sec) presence of ‘with the firewall with VPN’ is presented by triangle line. The average end-to-end delay slightly greater ‘with no firewall and no VPN’ in comparison ‘with Firewall no VPN’ and ‘with the firewall with VPN’ as all clients had willingly requested for all three applications like HTTP, FTP, Email so that’s why a huge amount of traffic was accessible. Besides, in the presence of a firewall, the firewall has blocked the HTTP traffic and the only VPN give open access to its users to use this traffic. When in the network, users are limited so its Average end-to-end delay will also get reduced like results shown below in Fig. 11. The graph shows the influence of the firewall and with a VPN on a cloud-based computing network. It is proved from the results that in presence of no firewall and no VPN average end-to-end delay was slightly greater than the presence of ‘firewall no VPN’ and ‘with the firewall with VPN’ in a network of cloud-based computing.

3) Average packet loss: Packet loss happens whenever a packet flops to the extent of the target while roaming through a network of computers. It is normally initiated by crowding over a network. In the presence of Firewall and VPN, it is also significant to investigate the average packet loss.

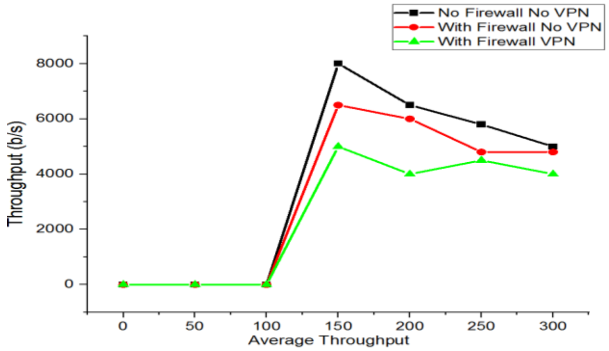


Fig. 10. Simulation Result of Average throughput for No Firewall No VPN, with Firewall No VPN and with Firewall VPN.

TABLE. III. RESULT OF AVERAGE END-TO-END DELAY FOR NO FIREWALL NO VPN, WITH FIREWALL NO VPN AND WITH FIREWALL VPN

Time	**Average End-to-end Delay (second)**		
	No Firewall No VPN	**Firewall No VPN**	**Firewall VPN**
0	0	0	0
50	0.00028745	0.000259	0.000259
100	0.000275809	0.000285	0.000285
150	0.000680813	0.000638	0.000621
200	0.000713229	0.000585	0.000659
250	0.000764209	0.000631	0.000712
300	0.000854009	0.000713	0.000738

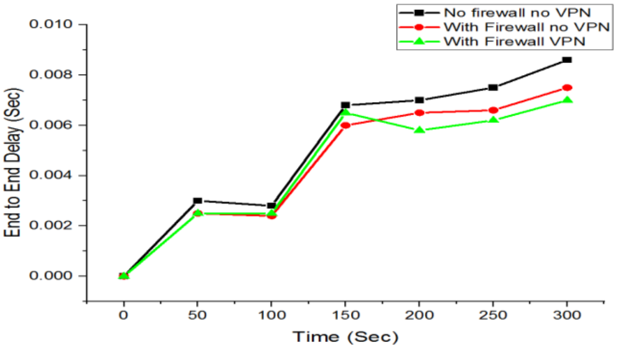


Fig. 11. Simulation Result of Average end-to-end Delay for No Firewall No VPN, with Firewall No VPN and with Firewall VPN.

Table IV represents the simulation results of average packet loss, simulation results of no Firewall no VPN, with Firewall no VPN and with Firewall VPN in the Cloud Computing network.

Fig. 12 demonstrates the comparison of average packet loss for ‘no firewall no VPN’, ‘with Firewall no VPN’ and ‘with a firewall with VPN’ in Cloud Computing. 18 nodes and 3 servers are included in the scenario. The simulation time per second is displayed on the horizontal axis, whereas the network average packet loss (packet loss/sec) is displayed on the vertical axis. The network average packet loss (packet loss/sec) presence of ‘no firewall and no VPN’ is represented by the square line while network average packet loss (packet loss/sec) presence of ‘with firewall no VPN’ is signified by circle line, whereas the network average packet loss (packet loss/sec) presence of ‘with firewall and with VPN’ is signified by triangle line. With the practice of ‘firewall and no VPN’, the average packet loss was greater in comparison with ‘no Firewall no VPN’ and ‘with firewall and with VPN’ because the firewall has blocked the traffic of Http and it is allowed for only users of VPN but all other clients can’t access traffic of Http. So, all packets of Http are dropped. The influence of firewall and VPN on cloud computing is displayed in results. It has been verified from the graph that average packet loss is greater in case of scenario ‘with firewall and no VPN’ as compared to the presence of ‘no firewall no VPN’ and ‘with firewall and VPN’ in a network of cloud-based computing. It is examined by wide simulation that firewall and VPN affect network performance of cloud though it provides better security.

B. *Average Http Traffic Comparison of No Firewall No VPN, with Firewall No VPN and with Firewall VPN*

Firewall blocked the traffic of Http while at VPN side Http traffic was simply allowable for users. Those users who are not using the service of the VPN tunnel cannot able to access Http traffic from the servers as the traffic was filtered by the firewall and tested that this request was of VPN. If those users were of VPN then they were allowable by a firewall to access the traffic of Http from the servers.

1) Server AA Average http traffic received: The server AA average Http traffic received the number of requests that were made by the user to server AA that was a part of a Cloud-based computing network. The server AA Http traffic received was presented in bytes per second (bytes/sec).

TABLE. IV. RESULT OF AVERAGE PACKET LOSS FOR NO FIREWALL NO VPN, WITH FIREWALL NO VPN AND WITH FIREWALL VPN

Time	Average Packet loss (Packet per second)		
	No Firewall No VPN	Firewall No VPN	Firewall VPN
0	0	0	0
50	1.94444	1.94444	1.94444
100	1.92156	1.94117	2.00000
150	2.01923	2.73202	2.48366
200	1.95588	3.16176	2.53921
250	1.97222	3.26190	2.47222
300	1.656667	2.93333	2.09000

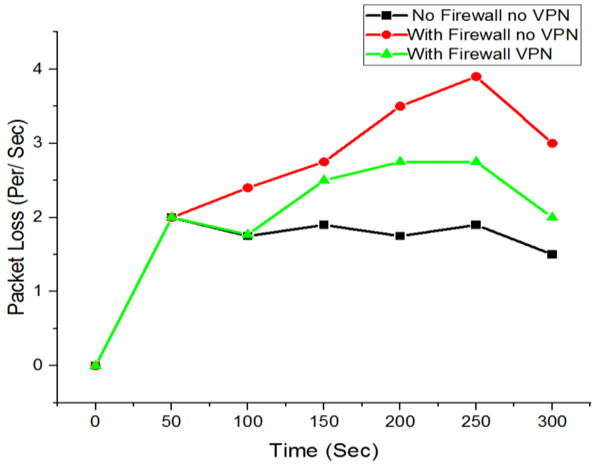


Fig. 12. Simulation Result of Average Packet Loss for No Firewall No VPN, with Firewall No VPN and with Firewall and VPN.

Table V defines the server AA average Http traffic received simulation results of no Firewall no VPN, with Firewall no VPN and with Firewall VPN

Fig. 13 illustrates the Server AA average Http traffic received simulation results of ‘no Firewall no VPN’, ‘with Firewall no VPN’ and ‘with Firewall and with VPN’ in a network of Cloud-based computing. 18 nodes and 3 servers are included in the scenario. Between those 18 nodes, the facility was not provided to 12 nodes to access Http traffic. Only those 6 nodes were VPN users that were connected to access point-1. These users were permitted to access Http traffic in the presence of firewalls and VPN. Simulation time per second is shown on the horizontal axis, though the Server AA average Http traffic received (bytes/sec) is displayed on the vertical axis. The Server AA average Http traffic received is displayed with the square line in the presence of ‘no firewall no VPN’ and the Server AA average Http traffic received is displayed with the help of circle line in the presence ‘with firewall no VPN’, while the Server AA average Http traffic received is shown with triangle line in the presence of firewall and VPN. In the presence of a VPN and firewall, the Server AA average Http traffic received was minimum in comparison with ‘no firewall and no VPN’, and the result shown in the graph is zero ‘with a firewall and no VPN’ as when the firewall is implemented for Http there are no facilities of VPN so no transmission took place between client and server. The results show the impact of firewall and VPN on Server AA in a network of Cloud-based computing. It has been revealed from the graph that the presence of ‘no firewall and no VPN’ give maximum Server AA average Http traffic received than the presence of ‘with Firewall no VPN’ and ‘with Firewall and with VPN’ in a network of cloud-based computing. Through extensive simulations It was showed that firewall and VPN affect the performance of the cloud however it gives better security.

2) Server AA Average http traffic sent: The server AA average Http traffic sent to represent the amount of data sent by the servers and received by the users which are present in the cloud computing network. The server AA average Http traffic sent is represented in bytes per second (bytes/sec).

TABLE. V. RESULT OF SERVER AA AVERAGE HTTP TRAFFIC RECEIVED FOR NO FIREWALL NO VPN, WITH FIREWALL NO VPN AND WITH FIREWALL VPN

Time	Server AA Average Http Traffic Received (bytes/sec)		
	No Firewall No VPN	Firewall No VPN	Firewall VPN
0	0	0	0
50	0	0	0
100	31.53153	0	18.91892
150	86.9281	0	64.05229
200	75.4902	0	49.03382
250	68.05556	0	40.2778
300	57.16667	0	33.83333

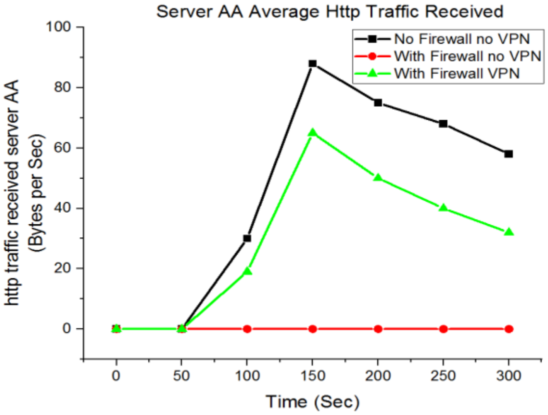


Fig. 13. Simulation Result of Server AA Average http Traffic Received for No Firewall No VPN, with Firewall No VPN and with Firewall VPN.

Table VI defines the server AA average Http traffic sent simulation results of no Firewall no VPN, with Firewall no VPN and with Firewall VPN in a network of Cloud.

Fig. 14 shows the server AA average Http traffic sent comparison of 'without Firewall and VPN', 'with Firewall and no VPN' and 'with Firewall and VPN' in the cloud computing network. 18 nodes and 3 servers are included in the scenario; among these 18 nodes, the 12 nodes do not have the accessibility to access Http traffic as just 6 nodes of VPN that were connected to access point-1 were the VPN users. They were only allowable to access Http traffic in the presence of a firewall with the help of VPN. The simulation time/second is showed on the horizontal axis, whereas the Server AA average Http traffic sent (bytes/sec) is shown on the vertical axis. The Server AA average Http traffic sent is displayed with the square line in the presence of 'no firewall no VPN' and the Server AA average Http traffic sent is presented with the circle line in the presence of 'firewall no VPN', whereas the Server AA average Http traffic sent is presented with triangle line in the presence of 'VPN and firewall'. The Server AA average Http traffic sent was minimum in the presence of a firewall and VPN as compared to 'no firewall and no VPN'. And the graph for 'with firewall and no VPN' is zero as whenever employing the firewall for Http 'without VPN so no Http communication was achieved between server and nodes. The results show the impact of firewall and VPN on server AA in a cloud computing network. From the graph, it has been verified that the presence of no firewall no VPN gives maximum server AA average Http traffic sent than the presence of 'with firewall no VPN and 'with firewall and VPN in a cloud computing network. Through extensive simulations, it was observed that firewall and VPN affect cloud performance while it gives better security.

TABLE. VI. RESULT OF SERVER AA AVERAGE HTTP TRAFFIC SENT FOR NO FIREWALL NO VPN, WITH FIREWALL NO VPN AND WITH FIREWALL VPN

Time	Server AA Average Http Traffic Sent (bytes/sec)		
	No Firewall No VPN	Firewall No VPN	Firewall VPN
0	0	0	0
50	0	0	0
100	18.65766	0	10.27027
150	58.88889	0	45.48366
200	51.60784	0	36.56373
250	46.3373	0	29.59921
300	38.92333	0	24.86333

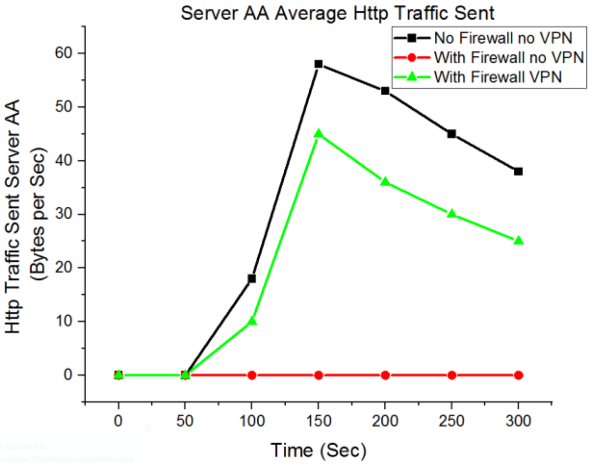


Fig. 14. Simulation Result of Server AA Average http Traffic Sent for No Firewall No VPN, with Firewall No VPN and with Firewall VPN.

V. CONCLUSION

In this paper, the research work links the VPN and Firewall effect on the performance of cloud computing. The cloud computing network is simulated and evaluated for without firewall and VPN with the help of OPNET modeler 14.5; and then compared and analyzed the performance of Cloud computing after deploying "with firewall and without VPN" and "with firewall and with VPN" in term of average throughput, average end-to-end delay and average packet loss. The simulation results indicated that average throughput and average end-to-end delay of the network was decreased when implementing firewall and VPN. It seemed from the results that IP VPN is a properly effective method for transferring of data over the Cloud computing network because it provides a suitable level of security and the end-to-end delay is unaffected in the network. Besides, simulation results also revealed the fact that the average packet loss increases with the presence of VPN and firewall. From the analysis, it is concluded that deploying the firewall and VPN slightly affects the performance of Cloud computing network while it gives better security.

REFERENCES

[1] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: state-of-the-art and research challenges," Journal of internet services and applications, vol. 1, pp. 7-18, 2010.

[2] A. Shawish and M. Salama, "Cloud computing: paradigms and technologies," in Inter-cooperative collective intelligence: Techniques and applications, ed: Springer, 2014, pp. 39-67.

[3] P. Mell and T. Grance, "The NIST definition of cloud computing," National institute of standards and technology, vol. 53, p. 50, 2009.

[4] R. Kumar, N. Gupta, S. Charu, K. Jain, and S. K. Jangir, "Open source solution for cloud computing platform using OpenStack," International Journal of Computer Science and Mobile Computing, vol. 3, pp. 89-98, 2014.

[5] A. E. Youssef, "Exploring cloud computing services and applications," Journal of Emerging Trends in Computing and Information Sciences, vol. 3, pp. 838-847, 2012.

[6] P. Sareen, "Cloud computing: types, architecture, applications, concerns, virtualization and role of it governance in the cloud," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, 2013.

[7] V. Chang, "Towards a Big Data system disaster recovery in a Private Cloud," Ad Hoc Networks, vol. 35, pp. 65-82, 2015.

[8] B. D. Cohen and B. L. Greenlaw, "Designing a Modern Software Engineering Training Program with Cloud Computing," 2017.

[9] S. Goyal, "Public vs private vs hybrid vs community-cloud computing: a critical review," International Journal of Computer Network and Information Security, vol. 6, p. 20, 2014.

[10] R. Buyya, S. N. Srirama, G. Casale, R. Calheiros, Y. Simmhan, B. Varghese, E. Gelenbe, B. Javadi, L. M. Vaquero, and M. A. Netto, "A manifesto for future generation cloud computing: research directions for the next decade," ACM computing surveys (CSUR), vol. 51, p. 105, 2018.

[11] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, and I. Stoica, "A view of cloud computing," Communications of the ACM, vol. 53, pp. 50-58, 2010.

[12] M. T. Khorshed, A. S. Ali, and S. A. Wasimi, "A survey on gaps, threat remediation challenges and some thoughts for proactive attack detection in cloud computing," Future generation computer systems, vol. 28, pp. 833-851, 2012.

[13] D. Teneyuca, "Internet cloud security: The illusion of inclusion," Information Security Technical Report, vol. 16, pp. 102-107, 2011.

[14] A. Joint, E. Baker, and E. Eccles, "Hey, you, get off of that cloud?," Computer Law & Security Review, vol. 25, pp. 270-274, 2009.

[15] Adil Khan and Jiang Feng, "Mobile Sink Random Mobility Model Impact in Wireless Sensor Nodes Energy Consumption Efficiency", International Review of Basic and Applied Sciences (IRBAS), Vol. 4, Issue 12, pp. 317-325, 2016.

[16] S. Y. Ameen and S. W. Nourildean, "Firewall and VPN investigation on cloud computing performance," International Journal of Computer Science and Engineering Survey, vol. 5, p. 15, 2014.

[17] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D. Zagorodnov, "The eucalyptus open-source cloud-computing system," in Cluster Computing and the Grid, 2009. CCGRID'09. 9th IEEE/ACM International Symposium on, 2009, pp. 124-131.

[18] B. Harris, "Firewalls and virtual private networks," 1998.

[19] P. Gupta and A. Verma, "Concept of VPN on cloud computing for elasticity by simple load balancing technique," International Journal of Engineering and Innovative Technology, pp. 274-278, 2012.

[20] A. Malik and H. K. Verma, "Performance Analysis of Virtual Private Network for Securing Voice and Video Traffic," International Journal of Computer Applications, vol. 46, pp. 25-30, 2012.

[21] H. Farman, B. Jan, M. Talha, A. Zar, H. Javed, M. Khan, A. U. Din, and K. Han, "Multicriteria-Based Location Privacy Preservation in Vehicular Ad Hoc Networks," Complexity, vol. 2018, 2018.

[22] H. Bourdoucen, A. Al Naamany, and A. Al Kalbani, "Impact of implementing VPN to secure wireless lan," World Academy of Science, Engineering and Technology, vol. 51, pp. 625-630, 2009.

[23] R. Buyya, S. N. Srirama, G. Casale, R. Calheiros, Y. Simmhan, B. Varghese, E. Gelenbe, B. Javadi, L. M. Vaquero, and M. A. Netto, "A manifesto for future generation cloud computing: Research directions for the next decade," ACM computing surveys (CSUR), vol. 51, pp. 1-38, 2018.

[24] K. Hamlen, M. Kantarcioglu, L. Khan, and B. Thuraisingham, "Security issues for cloud computing," International Journal of Information Security and Privacy (IJISP), vol. 4, pp. 36-48, 2010.

[25] A. R. Khakpour and A. X. Liu, "First step toward cloud-based firewalling," in 2012 IEEE 31st Symposium on Reliable Distributed Systems, 2012, pp. 41-50.

[26] F. Parkar and K. Wong, "Analysis of IP VPN Performance."

[27] S. Yu, R. Doss, W. Zhou, and S. Guo, "A general cloud firewall framework with dynamic resource allocation," in 2013 IEEE International Conference on Communications (ICC), 2013, pp. 1941-1945.

[28] J. Cropper, J. Ullrich, P. Frühwirt, and E. Weippl, "The role and security of firewalls in iaas cloud computing," in 2015 10th International Conference on Availability, Reliability and Security, 2015, pp. 70-79.

[29] Y. Yang, L. Wu, G. Yin, L. Li, and H. Zhao, "A survey on security and privacy issues in Internet-of-Things," IEEE Internet of Things Journal, vol. 4, pp. 1250-1258, 2017.

[30] G. Fylaktopoulos, M. Skolarikis, I. Papadopoulos, G. Goumas, A. Sotiropoulos, and I. Maglogiannis, "A distributed modular platform for the development of cloud-based applications," Future Generation Computer Systems, vol. 78, pp. 127-141, 2018.

[31] P. E. Idoga, M. Toycan, H. Nadiri, and E. Çelebi, "Assessing factors militating against the acceptance and successful implementation of a cloud-based health centre from the healthcare professionals' perspective: a survey of hospitals in Benue state, northcentral Nigeria," BMC medical informatics and decision making, vol. 19, p. 34, 2019.

AUTHORS' BIOGRAPHIES

Hussain Shah received MS degree as Gold medalist from the Institute of Computer Science & Information Technology, University of Agriculture Peshawar, KP, Pakistan. He is a PhD Scholar at the Department of Computer Sciences, Islamia College University Peshawar, KP, Pakistan. Currently, He is a Lecturer at the School of Computer Science, Shaykh Zayed Islamic Centre, University of Peshawar, KP, Pakistan. He is interested in Wireless Sensor Networks, Mobile Ad-hoc Networks, IoT, Cloud Computing and Image Processing.

Dr. Aziz-ud-Din received MS from University of Peshawar, KP, Pakistan, and PhD from UNIMAS Malaysia. He is currently working as an assistant professor at the School of Computer Science, Shaykh Zayed Islamic Centre, University of Peshawar, Peshawar, KP, Pakistan. He is interested in NLP, Mobile Ad-hoc Networks, IoT and Distributed system.

Abizar received MS degree from IBMS, Agriculture University of Peshawar. He is a PhD Scholar at the Department of Computer Sciences, Islamia College University Peshawar, KP, Pakistan. Currently, he is working as a Lecturer at the School of Computer Science, Shaykh Zayed Islamic Centre, University of Peshawar. His area of Interest includes WSN, Mobile Ad-hoc Networks, IoT and Smart Transportation. His Open researcher contribution identification (ORCID ID) is https://orcid.org/0000-0001-9472-6846.

Adil Khan received C.T. from AIOU Islamabad, B. Ed from the University of Peshawar, BS Honors in Computer Science from Edwards College Peshawar, M.S in Computer Science from City University of Science and Information Technology Peshawar and PhD from the University of Peshawar, Peshawar, KP Pakistan. He has over ten years of teaching, research and laboratory experience. In 2014-2016, he was a Senior Lecturer in Higher Education Department, Government of Khyber Pakhtunkhwa, Pakistan and in 2016-2019 he was a research scholar at the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001 PR China. Currently, Adil khan is working as an Assistant Professor at the School of Computer Science, SZIC, University of Peshawar. He has published many publications in top-tier academic journals and conferences. He is an Associate Editor and Reviewer for several journals. Adil Khan is interested in Cloud Computing, Machine Learning, Neural networks, Game Artificial Intelligence (Game-AI), Computer Vision, Image Processing (Breast Cancer Detection) and Social Network Analysis & Mining. His Open researcher contribution identification (ORCID ID) is http://orcid.org/0000-0003-2862-5718.

Shams ud Din received MS from Islamia College University Peshawar. He is pursuing a PhD at the Department of Computer Science, Islamia College University, Peshawar, Pakistan. Currently, he is working as a Lecturer at the same Department. His fields of interest include Cloud Computing, IoT, Computer Vision, digital image processing, and Machine Learning.

Cloud Computing Adoption at Higher Educational Institutions in the KSA for Sustainable Development

Ashraf Ali
Department of Computer Science and Information
College of Science Al Zulfi, Majmaah University
Kingdom of Saudi Arabia

***Abstract*—Rapid changes in the advancements of information and communication technologies (ICT) have prompted Higher Educational Institutions (HEIs) to enhance teaching and learning. Over the years, cloud computing (CC) has become an emerging and adoptable paradigm in many industries including healthcare, finance, and law with its promising benefits. This trend is also growing in the field of education around the globe. Due to its inherent qualities of reliability, scalability, flexibility and reasonable cost, cloud is the solution that addressed the accessibility issue for quality education. CC plays an important role and will have major impacts on HEIs of the Kingdom of Saudi Arabia (KSA) in the near future. HEIs are used to utilize the benefits of CC based services provided by the cloud service providers (CSPs). The CSP can be owned by the KSA government, private, or third-party vendors. By using cloud-based services at HEIs, staff, faculty, and students can utilize its services to perform various academic responsibilities on demand. This paper aims to adopt CC for HEIs and explore the prominent features and potential benefits of adopting cloud services in the HEIs of KSA. This paper also reveals numerous challenges, impacts, and major issues involved in adopting cloud services for HEIs.**

Keywords*—Cloud computing; higher educational institutions; Cloud Service Provers (CSP); Software-as-a-Service (SaaS); Platform-as-a-Service (PaaS); Infrastructure-as-a-Service (IaaS)

I. Introduction

In recent years, CC has emerged as a promising paradigm and received significant attention from many industries. CC is growing very fast and being adopted by many business domains in the world. According to the recent Gartner, Inc. report[1], the significant changes in the cloud market have been noticed. The worldwide public cloud market projected to grow very fast as compared to the previous years. It is predicted that the global cloud market is projected to grow from $ 182.4 in 2018 to $ 331.2 billion in the year 2022. The compound annual projected growth of the cloud market prediction is 18.42 %, which indicates significant changes in the coming years. Fig. 1 summarizes the "worldwide public cloud service revenue forecast" for PaaS, SaaS, and IaaS.

Recently, CC has sparked a revolution in higher education and has attained a major role in providing HEIs services. With its new promising benefits, CC enables all its HEIs stakeholders including students, academic staff, administrative staff, and other key stakeholders to access services that are hosted on cloud (SaaS, PaaS, IaaS) [1],[2] on demand. Nowadays, HEIs in many countries have started to implement cloud models. There are 3-types of cloud models (Private, Government, and Third-Party Clouds). Some universities in the world started to implement a private cloud, which have to establish self-owned cloud [3], some Universities use government-owned shared clouds [4], and some Universities buy and avail third-party vendors [5] cloud services.

This paper aims to identify and explore the promising benefits and challenges of CC adoption at HEIs in the KSA. The author explores major promising benefits including cost-effectiveness, high capability, greater flexibility, 24/7 accessibility, openness, boost collaborative work, accountability, great opportunities with a variety of choices in teaching, learning, management, and challenges including security, integrity, availability, improper management, training for staff, complexity, and affordability. The author also discusses the current state of the cloud adoption at HEIs in KSA. This research will help to motivate the HEIs of KSA to adopt cloud-based services for sustainable development and to achieve one of its promising goal of KSA Vision 2030.

The remainder of this paper organized as follows. Section II describes CC, its characteristics, and CC at HEIs. Section III describes stages in the adoption of CC services at HEIs. Section IV discusses the present state of cloud adoption of KSA Universities. Section V explores the CC adoption advantages, issues, and challenges at HEIs. Finally, Section VI concludes the paper.

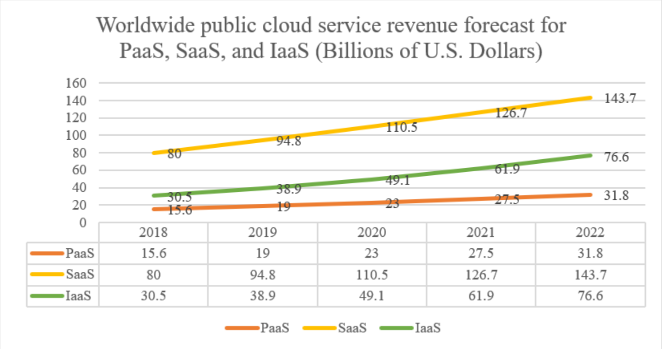


Fig. 1. Worldwide Public cloud Service Prediction (Source: Gartner Inc, April 2019).

[1] https://www.gartner.com/en/newsroom/press-releases/2019-04-02-gartner-forecasts-worldwide-public-cloud-revenue-to-g

II. Cloud Computing

CC is an idea of combining technological resources online. This term first introduced by Google in 2006 and has become the standard in the modern days computing environment. The online resources (Networks, servers, storage, and applications, etc.) can put in the common pool where any individual or company can pay-per-use the series of services according to their specific needs with minimal management efforts. Cloud leverages several elements including scale, cost-efficiency, resilience, service orientation, agility, etc. This pay-per-use feature of cloud models enables the user to access the resources as per their requirement from the shared-pool online resources available on the network. The end-user can access resources 24/7 from anywhere using various types of devices such as laptops, desktop, and smartphone. The platforms (System Software) and infrastructure (Hardware) used to execute various application software which user access and use online on-demand (Fig. 2).

According to the NIST CC definition [25], there are "five essential characteristics", "three service models", and "four deployment models" as given in Fig. 3.

CC definition models are interrelated and work together to perform the desired task as per its characteristics. CC definition model is shown in Fig. 3 and demonstrates how each of the five essential characteristics of CC is encompassed as an integral part of each cloud service model (IaaS, PaaS, and SaaS). According to the definition determined by the NIST [25] to provide an IaaS, PaaS, and SaaS solution for cloud services. These include all five characteristics of CC. If it does not satisfy any one of the characteristics, then it is not offering real cloud service as per NIST definition [25]. The creation of four-cloud deployment models (Private, Public, and Community clouds) is the result of integrating any one or more service models.

A. CC Characteristics

CC is different from outsourcing software services. It has five key characteristics that distinguish from outsourcing software services [6]. These characteristics of CC are given in Fig. 4.

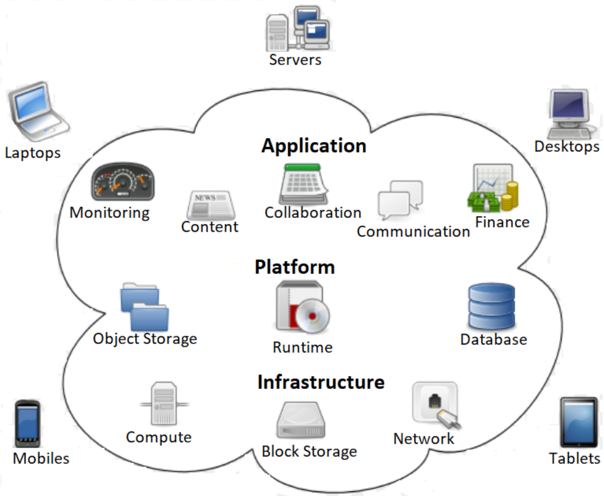


Fig. 2. Cloud Computing.

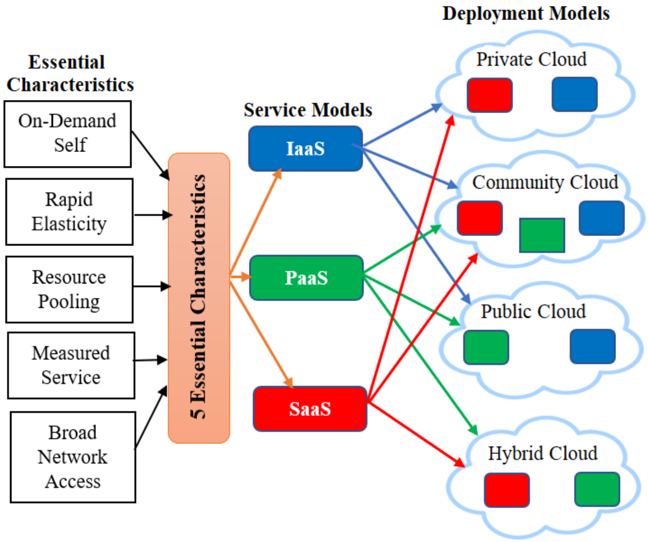


Fig. 3. NIST CC Definition Models [25].

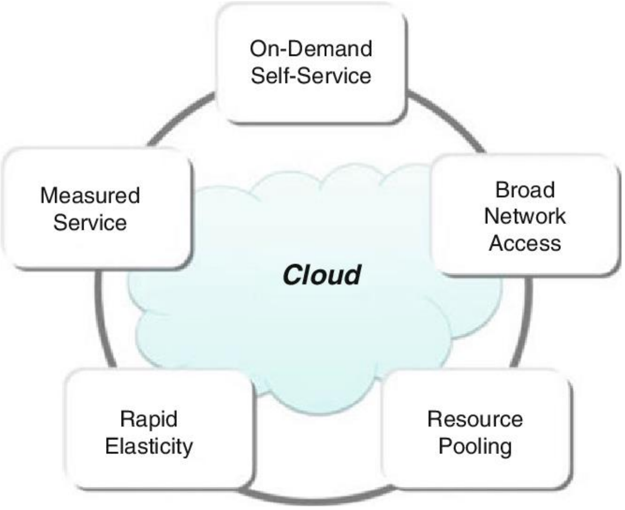


Fig. 4. Five Key Characteristics of CC [6].

These characteristics can be utilized by HEIs as follows:

1) Broad Network Access: Stakeholders of HEIs such as students, academic staff, and other key stakeholders can access network resources by using various devices.

2) Measured Service: This service allows the HEIs key stakeholders to automatically control and improve the use of resources by introducing measured services which can be later modified as per the requirement of HEIs. The key characteristic of this feature is to provide an efficient, fruitful, and cost-effective service. This enables HEIs stakeholders to use the cloud services "as pay and use service" without wasting money on downtime.

3) On-demand Self-Service: This service allows the HEIs stakeholders to access a variety of resources (email, storage, and applications, etc.) from anywhere at any time.

4) Rapid Elasticity: This service enables the HEIs stakeholders to process, utilize and adjust the cloud resources to meet the requirements according to their demand.

5) Resource Pooling: the main objective of this service is to enable the HEIs stakeholders to use pooled cloud resources via networks according to their demands.

The aforementioned characteristics related to the major advantages of cloud services in regards to HEIs, which motivate to adopt CC for efficient, fruitful, cost-effective services. Cloud services permit HEIs to develop cost-effective and quality education at a global level. Shyan [7], studied cloud services and found that we cannot give financial justification when one is dependent on an individual machine to fulfill the computing requirements. Seigle [8], found that combining the various IT services to justify the quality of service and efficient nature of the cloud. The advancements of modern technologies, cost-reduction, flexibility, and quality aspects of the cloud are a clear justification for the CC adoption by HEIs [2], [9]. Therefore, CC comprehends a paradigm shift for ICT advancement in HEIs.

B. *CC at HEIs*

HEIs play a vital role in building better societies. The education field has always incorporated new technologies and teaching methodologies towards educational empowerment. Over the years, most of the HEIs in the world are becoming extremely dependent on modern ICT in terms of fulfilling their service requirements for content delivery, management, communications, and collaboration. These services are increasing very fast. These services provided using high-speed internet and web browsers access to the HEIs key stakeholders including faculty members, students, administrative staff, and other members.

In the traditional system, HEIs requires to spend huge costs for the ICT infrastructure. It also requires having a highly skilled IT Department and many demanding software solutions. It requires spending lots of time and money. Over the last two decades, CC has empowered HEIs to offer a variety of on-demand cost-effective services efficiently. These cost-effective services are used by HEIs key stakeholders to support learning, accessing online classes, online class registration, social interaction, content creation, course design, and class preparation daily. For example, cloud-based services include OneDrive, Dropbox, and Google Apps, Social media such as Facebook, YouTube, and Twitter. HEIs cloud can provide a variety of services as shown in Fig. 5.

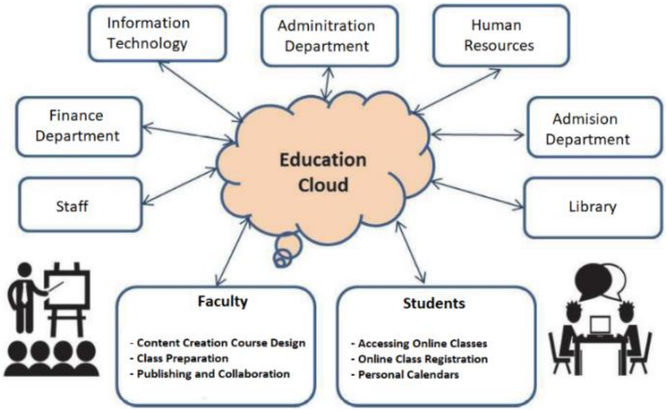


Fig. 5. Cloud and its Services at HEIs.

TABLE I. COMMON CLOUD SERVICE MODEL AND KEY USERS AT HEIS.

Service Models	***Descriptions***	***Key Users***
SaaS	SaaS entirely depends on the internet where applications are deployed and available on-demand. End-users can access applications using the browser or by using an interface. Subsequently, application services offered to the end-user on-demand through software. This feature enabled end-user to deploy the services quickly, which brings ease of use and monetary benefits. Many CSP corporations in the world offering these types of platforms such as GoogleApps (email, google docs, and calendar), Salesforce, NetApp, ServiceNow, and GotoMeeting.	Faculty, Staff, Students, classes and Admin Department
IaaS	This is a virtual platform that allows end-users to deploy their applications on cloud infrastructure and use services such as networking, storage, database, backup, security, and other resources. This platform gives full control to the user over operating systems and deployed applications. Many companies in the world provide these types of platforms such as AWS, Rackspace, AttendaRTI, Amazon EC2, and CenturyLink	Servers, Data Storage, IT Department, and Researchers
PaaS	This is a platform-based service having a suite of cloud services that provides end-users to develop, execute, manage, and integrate on-premises and cloud-based applications and services. This feature offers no waiting time required for the deployment of appropriate applications (hardware and software). End-user can build applications by utilizing the platform and using the tools, languages, and services supported by CSP. Many companies in the world provide these types of platforms such as GoogleAppEngine, Azzure, AWS, OCP, and Salesforce.	Execution, Database, Researchers, and Developers

Cloud services delivered through different service model structures. Table I [10] summarized the most common cloud service models and their key users at HEIs.

Ali MB et al. [11], have recently reviewed various papers related to cloud adoption in HEIs. Recently, they have studied 20 research papers related to CC adoption in HEIs. From their study, they have examined the major advantages, challenges and several issues towards cloud adoption in HEIs. They confirmed that there is an increasing interest from various HEIs around the globe to migrate to the cloud.

III. STAGES IN THE ADOPTION OF CLOUD SERVICES AT HEIS

The adoption of cloud services at HEIs requires a well-defined cloud strategy that supports CC capabilities. For a well-defined cloud strategy, there is a need to hire experts from various fields. This will help the administrators to fully understand the benefits of the cloud-based services model, its impact and the major issue of CC adoption. The strategy implementation of cloud services involves creating a new framework that must be built according to the need of key stakeholders such as faculty, students', and board directors of the HEIs. To have a successful cloud strategy, the key stakeholder should primarily participate to define the HEIs cloud strategy that addresses its opportunities, challenges, and issues specific to HEIs as well as cloud strategy must be aligned with HEIs strategy. Many HEIs in the world have not been successful in the adoption of cloud service because of the failure of cloud-strategy. Moreover, many HEIs around the globe do not know how to take initiative to start cloud projects.

A. Stage 1 (Identify Cloud Service Requirement)

This step requires HEIs to intricately study the cloud and achieve a comprehensive analysis of the applications and service requirements. From the software perspective, it also needs to determine the feasibility of the project. The cloud services delivery to the HEIs key stakeholders should be based on their requirements. End-user services are offered on-demand and are billed pay-per-use. The requirements and feasibility study need to be determined in this stage, and a crucial plan to be developed. This stage also elaborates on how the strength and opportunities of the existing system can be maximized and weaknesses and threats can be minimized. Furthermore, this stage will be determining what services are needed and how much services will have to be consumed for the faculty, staff, students and other HEIs key stakeholders.

B. Stage 2 (Evaluation Stage)

This step needs HEIs to study and analyze the current stage of the IT internal process and its impact on cloud services adoption. This step also elaborates on the experimental use of cloud services, opportunities, challenges, and major issues. The traditional system needs to be reviewed and all the service conveyance needs to be considered. This stage also stressed the demand for refinement of the traditional system and the elimination of inefficient processes. Security, legal issues, and other issues which are identified in stage 1 need to be set. HEIs' finest policies, practices, and how these can be achieved when shifting to the cloud. This stage also considers the broader impacts on HEIs such information and data security protection and supports services. The main objective of this stage is to identify emerging technologies and their efficiency in terms of cost and time [12].

C. Stage 3(Choosing Solution and Migration)

This step requires to choose the solution for the actual migration of the systems. End-user should go through in-depth study for the internal IT processes with their offering services to determine the overall architecture of the workload, the security profile for each workload, and its local environment. In this stage system integration needs to be done after the conclusion. In conclusion, the decision should be made for the cloud deployment model. HEIs must decide workload and its platform, run on either public cloud, private cloud or combination of both. Application and data migration can proceed in this stage. Support services should be provided to the end-user during the migration. Monitoring and control for the project needed to ensure a successful migration. All the practices, learning lessons, cost management should be documented and must be communicated to all HEIs key stakeholders. For HEIs, investments in migration to cloud-based services should be based on maximizing the current resources that they already have such as networking, database, and storage. The analysis of the cost-effectiveness, high performance, and network connectivity must also be taken into account. The primary benefits and significant challenges must be stressed in providing short-term and long-term storage in the cloud. Criteria need to be finalized to find out the best cloud solutions that offer scalable, reliable, secure, and flexibility. For example, data security, rapid development, and flexibility for providing facilities in e-Learning and distance education should be focused on.

IV. Present State of Cloud Adoption of KSA Universities

M. Al-Ruithe et al. [24], recently reviewed the present state of cloud adoption in the public sectors of KSA. From their study they found that the majority of public sectors (54.37 %) have not adopted cloud-based services in their organizations, 29.13 % adopted some form of cloud-based services, and 16.50 % don't know about it. To do the measurement of the present state of cloud adoption at HEIs in KSA, the author has followed the same approach. The author has surveyed to make an analysis of the present state of cloud adoption and cost comparison among traditional and cloud-based services at Universities in KSA. Questionnaires were distributed to various Universities in KSA to the members of the various departments including Administrative, Academic, and IT. The questionnaire has been distributed to 90 correspondents, in which 71 responded. The response rate was 78.89 % which is considered a very good response. The first measurement was conducted to find out the cloud adoption with the direct response ("Yes"/ "No"/ "Don't know"). The survey shows that the majority of the Universities in KSA (52.12 %) still have not adopted cloud services. 16.89 % responded that they "Don't know" and 30.99 % responded that they have already adopted some cloud-based services especially SaaS in their Universities (Fig. 6).

For those who reported with "No", the data has been collected with the direct sub-questions ("intend to adopt in the next 1 year"/ "intend to adopt in the next 2 years"/ "not yet decided"/ "don't know"). 16.22 % reported they "intend to adopt in the next 1 year", 37.84 % reported they "intend to adopt in the next 2 years", 32.43 % reported they have "not yet decided", and 13.51 % reported they "don't know" about it (Fig. 7).

This section described who have already adopted CC models and reported consuming cloud services. SaaS service model adopted by 54.55 %, while IaaS adoption rate is 31.82 % and PaaS adoption rate is 13.63 % (Fig. 8).

The KSA Universities' organizational structure is shown in Fig. 9. As can be seen from the KSA Universities' organizational structure there are many offices including vice-rectors', finance, deanships, and colleges working under the Rectors' office. Different universities have various colleges and deanships. On average each University has 13 colleges. The number of departments at colleges also varies from college to college ranging from 4 departments to more.

To make analysis and do the comparison, data was collected in regards to calculating the approximate manual, electronic, and cloud-based services cost. Approximate cost estimation has been summarized in Table II. Due to the confidentiality and secrecy of the organization and department data has been compiled and the actual cost has been hidden.

Unit approximate cost for the particular offices, colleges, deanship, and others has been considered to do the analysis and comparison. In this study, the author has considered the number of units for Rector's office (1), Rector's Admin offices (5), Vice Rector's offices (4), Vice Rector's Admin offices (3), Colleges (13), and others (10).

From this study, it is confirmed that there will be a 73.97 % cost reduction as compared to the manual cost and 61.65 % cost reduction as compared to electronic cost. The study motivates HEIs to adopt CC based services due to the enormous cost reduction and prominent benefits. The study also elaborates adopting CC based services at HEIs in KSA will have great positive impacts. Fig. 10 represents the cost comparison analysis of traditional vs. cloud at KSA Universities.

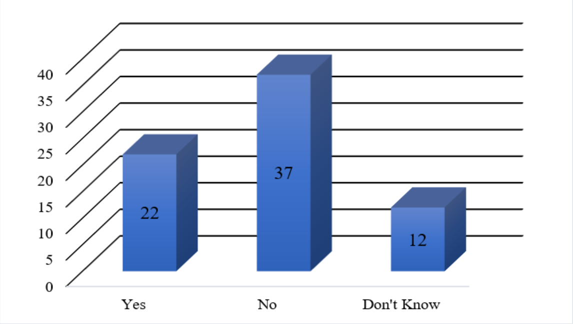


Fig. 6. The Present State of Cloud Adoption at KSA Universities.

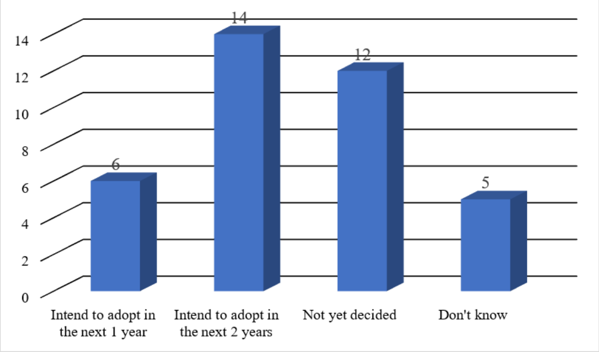


Fig. 7. KSA Universities Plan for cloud Adoption.

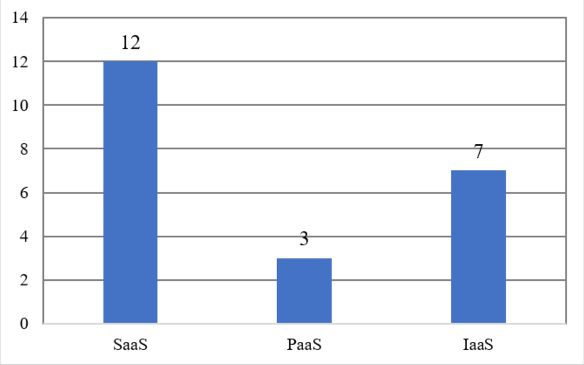


Fig. 8. Cloud Service Model Adopted by KSA Universities.

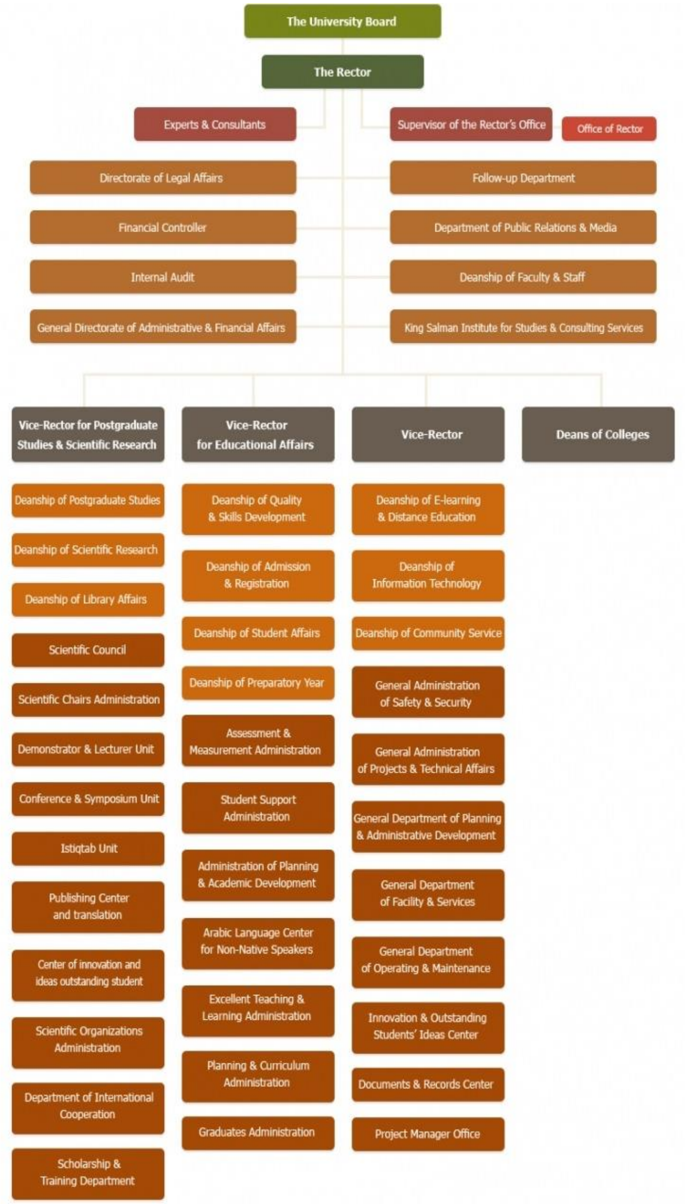

The University Board
The Rector
Experts & Consultants
Supervisor of the Rector's Office
Office of Rector
Directorate of Legal Affairs
Follow-up Department
Financial Controller
Department of Public Relations & Media
Internal Audit
Deanship of Faculty & Staff
General Directorate of Administrative & Financial Affairs
King Salman Institute for Studies & Consulting Services
Vice-Rector for Postgraduate Studies & Scientific Research
Vice-Rector for Educational Affairs
Vice-Rector
Deans of Colleges
Deanship of Postgraduate Studies
Deanship of Scientific Research
Deanship of Library Affairs
Scientific Council
Scientific Chairs Administration
Demonstrator & Lecturer Unit
Conference & Symposium Unit
Istiqtab Unit
Publishing Center and translation
Center of innovation and ideas outstanding student
Scientific Organizations Administration
Department of International Cooperation
Scholarship & Training Department
Deanship of Quality & Skills Development
Deanship of Admission & Registration
Deanship of Student Affairs
Deanship of Preparatory Year
Assessment & Measurement Administration
Student Support Administration
Administration of Planning & Academic Development
Arabic Language Center for Non-Native Speakers
Excellent Teaching & Learning Administration
Planning & Curriculum Administration
Graduates Administration
Deanship of E-learning & Distance Education
Deanship of Information Technology
Deanship of Community Service
General Administration of Safety & Security
General Administration of Projects & Technical Affairs
General Department of Planning & Administrative Development
General Department of Facility & Services
General Department of Operating & Maintenance
Innovation & Outstanding Students' Ideas Center
Documents & Records Center
Project Manager Office

Fig. 9. The Organizational Structure of KSA Universities (Source: Majmaah University, KSA).

TABLE II. APPROXIMATE COST ESTIMATION AMONG MANUAL, ELECTRONIC AND CLOUD-BASED SERVICES

University Offices	***Manual Cost***	***Electronic Cost***	***Cloud Cost***
Rector's office	12	10	2
Rector's Admin Offices	25	15	5
Vice Rector's offices	16	8	4
Vice Rector's Admin Offices	30	21	6
Colleges	39	26	13
Deanship's	44	33	11
Others	30	20	10

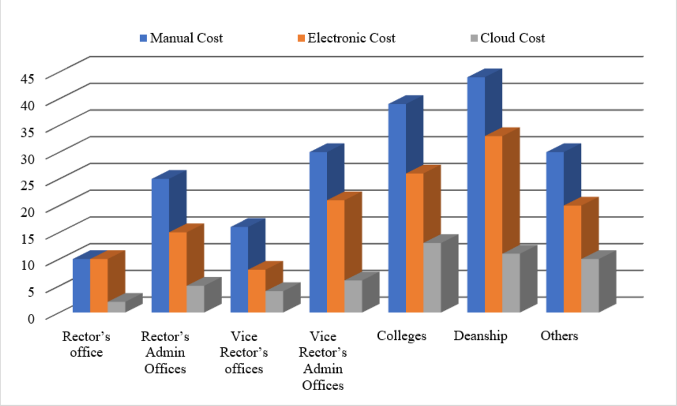


Fig. 10. Traditional Vs. Cloud cost comparison at KSA Universities.

V. CC Adoption Advantages, Issues, and Challenges at HEIs

The major objective of adopting CC at HEIs is to save both time and money. CC can eradicate the requirement of various specialized software and hardware on individual faculty or departments of the HEIs. This, in turn, allows for a greater degree of flexibility to access and scales various hardware and software resources. Any individual member or department will be able to access the required information stored on the "Cloud" from anywhere they have access to the internet. This will allow every individual to access and submit their information from anywhere at any time. For example, professors can submit the grade of students while they are out of the office and students to be able to access their assignments, lecture notes and emails during the semester breaks. Over the years, several researchers described many benefits of CC in HEIs. CC generally employed in HEIs for a student information system (SIS) and Learning Management Systems (LMS) [13]-[16].

CC adoption at HEIs will have a great impact on the reduction of capital investment in software and hardware infrastructure [11], [14]. Over the years many researchers have explored promising benefits of cloud-based services such as cost-effectiveness, greater flexibility, utilization of resources, rapid development, easy maintenance, and increased scalability, enhance availability, [2], [11], [14], [17]-[19],[23]. Cloud-based services provide great opportunities for HEIs key stakeholders with a variety of choices in teaching, learning, and management. This enables faculty, students, and staff to use many applications free of cost, or pay-per-use. These features also enable HEIs to focus more on their research and teaching goals. Cloud-based services offer openness to students and other stakeholders to use new technologies. Cloud-based services also offer more flexibility to faculty, students, and staff to acquire resources and use services on-demand. Cloud-based services boost collaborative work, which enables students, faculty, and staff to access resources and applications on-demand from their computers without the installment of specific software. This facility enables flexibility and facilitates interdepartmental collaboration. Every HEIs need to submit Quality Accreditation Report after a specific period. It is a continuous process. To submit the report there is a need to provide lots of evidence for quality accreditation. Cloud-based services improve accountability and enable the HEIs stakeholder to collect a large amount of information just by submitting them into the system without wasting time.

In addition to the promising advantages, cloud-based services include various security issues and challenges in CC adoption at HEIs [2],[11],[20]-[22]. These issues and challenges must be managed for a truly successful paradigm shift for the HEIs. The major issue is the lack of control and ownership of data. In CC, data and information are not stored on-premises of HEIs and data accessed through the internet. The third-party provides these services. It is the responsibility of the HEIs to ensure that it implements all security features to protect information against unauthorized access, modification or destruction of data and information. CSPs must provide appropriate standards to protect the CC system and ensure the confidentiality, availability, and integrity of the system. To ensure these HEIs may require using a strong encryption mechanism, and other security features to protect data and information. In addition to security issues, there are some challenges at HEIs such as improper management, training for the staff, integration, complexity, and affordability. Lack of training to the staff in the implementation or use of the cloud can lead to the inadequate utilization of the cloud. For example, a virtual lab left open and not managed properly will increase the cost irrespective of actual usage. Proper cloud management will have a great impact on effective consumption, operation, and optimization of cloud services. Table III summarizes the advantages, issues, and challenges at HEIs.

TABLE III. CC Adoption Advantages, Security Issues and Challenges at HEIs

✓ **Advantages**	✓ **Security Issues**	✓ **HEIs Challenges**
✓ Cost Reduction ✓ Access to applications from anywhere ✓ Centralized Management ✓ Boost collaborative work ✓ Rapid development and increased scalability ✓ Easy customization, continuous improvements. ✓ Free software or pay-per-use ✓ Support for teaching and learning ✓ Enhanced 24/7 availability ✓ Reallocation resources and easy maintenance ✓ Increased openness of students and other stakeholders to new technologies. ✓ Improves accountability and automatic provisioning ✓ Increased utilization through sharing of the resources ✓ More flexibility: acquire resources and use services on-demand	✓ Information Security ✓ Data Protection ✓ Data privacy ✓ Data Sanitization ✓ Unauthorized Access ✓ Lack of control ✓ Confidentiality ✓ Integrity & Availability	✓ Improper Management ✓ Training for staff ✓ Integration ✓ Inadequate computers ✓ Poor internal access ✓ Culture of HEIs ✓ Complexity and Affordability

VI. Conclusion

The KSA government "Cloud First" policy[2] is an effective initiative in the reforming process to achieve the goal of its promising KSA vision 2030 under the e-University theme. CC is now a mainstream paradigm of IT services that recognized as a massive benefit to HEIs. In recent years, cloud usage in HEIs around the globe is common and becoming widely recognized in the field of education. CC enables all its HEIs stakeholders including students, academic staff, administrative staff, and other key stakeholders to access services by using any device anywhere at any time. Many HEIs in the world have already adopted cloud services for SIS and LMS. In this paper, the author explores the promising benefits of cloud-based services including cost-effectiveness, high capability, greater flexibility, 24/7 accessibility, openness, boost collaborative work, accountability, great opportunities with a variety of choices in teaching, learning, and management. All of these are contributing to the rapid acceptance of CC in HEIs for sustainable development. Issues and challenges must be understood, managed for a true paradigm shift for HEIs. Security, integrity, availability, improper management, training for staff, complexity, and affordability are some of the major implementation challenges that need to be taken into account. The author also discusses the present state of cloud adoption at HEIs in KSA. Around 52% of KSA Universities presently do not adopt cloud services. cloud-based services adoption is a foremost decision for any HEIs. The main contribution of this work is to motivate the HEIs of KSA to adopt cloud-based services for sustainable development and to achieve one of its promising goal of KSA Vision 2030.

Acknowledgment

The author would like to thank the Deanship of Scientific Research at Majmaah University for supporting this work under project no. R-1441-97.

References

[1] D. Manzoor, A. Ali, and A. Ahmad, “Cloud and Web Technologies:Technical Improvements and Their Implications on E-Governance,” International Journal of Advanced Computer Science and Applications, vol. 5, no. 11, pp. 196–201, May 2014.

[2] N. Sultan, “ Cloud computing for education: A new dawn?,” International Journal of Information Management, vol. 30, no. 2, pp.109-116. April 2010.

[3] HE. Schaffer et. al., “NCSU’s virtual computing lab: A cloud computing solution,” Computer, vol. 42, no. 7, pp. 94-97, July 2009.

[4] The determinants of cloud computing adoption by colleges and universities, WE. Klug, USA: Northcentral University, 2014.

[5] AA. Shakeabubakor, E. Sundararajan, and A. R. Hamdan, “Cloud computing services and applications to improve productivity of University researchers," International Journal of Information and Electronics Engineering, vol. 5, no. 2, March 2015.

[6] F. Shahzad, “State-of-the-art survey on cloud computing security Challenges, approaches, and solutions," Procedia Computer Science, vol. 37, January 2014.

[7] J. Shayan et. al., “Identifying Benefits and risks associated with utilizing cloud computing," arXiv preprint arXiv:1401.5155, January 2014.

[8] D. Siegle, “Technology: Cloud Computing: A Free Technology Option to Promote Collaborative Learning," Gifted Child Today, vol. 33, no. 4, pp. 41-45, October 2010.

[9] N. Sultan, “Reaching for the cloud: How SMEs can manage," International journal of information management, vol. 31, no. 3, pp. 272-278, June 2011.

[10] R. Matsumoto, “SaaS does not necessarily equal cloud," http://www.rickmatsumoto.com/saas-does-not-necessarily-equal-cloud/, 2012.

[11] MB. Ali, H. T.Wood, and M. Mohamad, “Benefits and challenges of cloud computing adoption and usage in higher education: a systematic literature review," International Journal of Enterprise Information Systems, vol. 14, no. 4, pp. 64-77, October 2018.

[12] C. Erenben, “Cloud computing: the economic imperative," ESchool News, vol. 12, no. 3, pp. 9-13, March 2009.

[13] AS. Noor, M. Younas, and M. Arshad, “A review on cloud-based knowledge management in higher education institutions," International Journal of Electrical and Computer Engineering, vol. 9, pp. 2088-8708, December 2019.

[14] M. Attaran, S. Attaran, and B. G. Celik, “Promises and challenges of cloud computing in higher education: a practical guide for implementation,"Journal of Higher Education Theory and Practice, vol. 17, no. 6, November 2017.

[15] A. Lin, and NC. Chen, “Cloud computing as an innovation: Perception, attitude, and adoption," International Journal of Information Management, vol 32, no. 6, pp. 533-540, December 2012.

[16] C. Boja, P. Pocatilu, and C. Toma, “The economics of cloud computing on educational services," Procedia-Social and Behavioral Sciences, vol. 93, pp. 1050-1054, October 2013.

[17] KE. Krelja, J. Tomljanovi?, and K. Broni?, “Usage of cloud applications by students," Zbornik VeleuiliŽta u Rijeci, vol. 2, no. 1, pp. 13-26, July 2014.

[18] J. Scholten, “The determinants of cloud computing adoption in The Netherlands: a TOE-perspective," Master’s thesis, University of Twente, 2017.

[19] C. Stergiou et. al., “Secure integration of IoT and cloud computing," Future Generation Computer Systems, vol. 78, pp. 964-975, January 2018.

[20] A. Haider A, “Business technologies in contemporary organizations: adoption, assimilation, and institutionalization: adoption, assimilation, and institutionalization," IGI Global, October 2014.

[21] MT. Amron, R. Ibrahim, and S. Chuprat, “A Review on Cloud Computing Acceptance Factors," Procedia Computer Science, vol. 124, pp. 639-646, January 2017.

[22] A. Alharthi et. al., “An overview of cloud services adoption challenges in higher education institutions," pp. 102-109, 2015.

[23] ZM. Jawad, IK. Ajlan, ZD. Abdulameer, “Cloud Computing Adoption by Higher Education Institutions of Iraq: An Empirical Study," Journal of Education College Wasit University, vol. 1, no. 28, pp. 591-608 August 2017.

[24] M. Al-Ruithe, E. Benkhelifa, and K. Hameed, “Current State of Cloud Computing Adoption-An Empirical Study in Major Public Sector Organizations of Saudi Arabia (KSA)," Procedia Computer Science, no. 110, pp. 378-85, Jan 2017.

[25] M. Peter, and T. Grance, “The NIST definition of cloud computing", 2011.

[2] https://www.mcit.gov.sa/sites/default/files/ksa_cloud_first_policy_en.pdf

Automatic Assessment of Performance of Hospitals using Subjective Opinions for Sentiment Classification

Muhammad Badruddin Khan
Information Systems Department, College of Computer and Information Sciences
Al Imam Mohammad ibn Saud Islamic University (IMSIU)
Riyadh, KSA

***Abstract*—Social media is the venue where the opinions are shared in form of text, images and videos by public. Hospitals' performance can be judged by opinions that are written by patients or their relatives. Machine learning techniques can be used to detect sentiments of the opinion givers. For the research work presented in this article, opinions for few big hospitals were collected using Facebook, twitter and hospitals' webpage. The corpus was constructed and the sentiment analysis was performed after few preprocessing tasks. Resources like Stanford POS Tagger and WordNet were used to discover aspects. In this paper, the challenges of annotation of subjective opinions are discussed in detail. Two sentiment lexicons namely NRC-Affect-Intensity lexicon and SentiWordNet 3.0 lexicon were used to calculate sentiment scores of the comments that were used by different machine learning classifiers. Moreover, the results of the experiments on the constructed dataset are provided. For the experiments that aimed to discover overall sentiments of user towards hospital, Random forest outperformed other classifiers achieving accuracy of 76.49% using scores from NRC-Affect-Intensity lexicon. For the experiments that were directed towards discovering sentiments of users towards particular aspect of a hospital, Random forest overtook other classifiers reaching accuracy of 80.7339 % using NRC-Affect-Intensity lexicon sentiment scores. The research results show that machine learning can be very helpful in identifying sentiments of users from their textual comments that are vastly available on different social media platforms. The results can be helpful in improvement of hospital performance and are expected to contribute to growing field of health informatics.**

Keywords*—*Health informatics; Classification Algorithms; Sentiment Analysis; Sentiment Lexicons; Text Mining

I. Introduction

There are numerous social networking websites such as, Facebook, Google Plus, Twitter, LinkedIn and etc. that have used information technology to contract this globe into a village. People connect to each other and share their opinions, emotions and sentiments in the form of posts and comments using various social networking websites. These posts and comments are valuable source of data that is growing at unprecedented rate. This huge data contains lot of hidden insights, which needs application of data/text mining techniques to be revealed. Education and health are the two most important sectors for the society. A person's deterioration of health affects entire family. Hospitals are the places where patients come with expectations to restore their health. The services provided by hospitals become part of their experiences. Social media is one of the means to make these good/bad experiences visible to the world. The experiences are shared in different forms. One can write blog post(s), share picture(s)/video(s) and compose comment(s) and these shared opinions act as a trigger and attract more people to share their own personal experiences. These personal experiences can be helpful and beneficial to hospitals' administration and based on opinions of their patients, they can take steps to improve different aspects of their hospitals. Moreover, those patients who plan to receive services of particular hospital in future, can see reviews from previous patients of that particular hospital and decide whether to go to such hospital or not. Machine learning algorithms can be helpful for the task of automatic analysis of such opinions and reviews.

In this paper, personal experiences shared in the form of posts and comments, were used to determine sentiments of people who received medical services from the hospital. Text mining techniques and different sentiment lexicons were used to discover sentiments of experience sharers and opinion givers. The positive side of involvement of machine learning technique to accomplish the task is that machines are expected to be unbiased and unbiased discovery of sentiments can be a useful asset for hospitals to understand their current situations and improve their future performances.

The text of opinions or comments that are shared on social media is not simple. Sometimes, it is even difficult for humans to understand the correct meaning of the comment. Moreover, the granularity level of sentiment in a comment also varies. It means that the text of comment does not always talk about the overall performance of the hospitals with sentences like: “This hospital is good” and “That hospital is not good”. The opinion-sharer or commenter can share his/her sentiment about particular aspect of a hospital. It is possible that a hospital performs well with respect to one performance criterion, but with respect to another criterion/criteria, people are not happy with it. These criteria and sentiments related to them are needed to be identified automatically. Aspect based sentiment analysis seeks to understand sentiments about different aspects for specific entities. In this work, the entities are the hospitals and aspects are their different performance criteria.

In this work, comments regarding performances of few big non-Government hospitals of Pakistan were collected. The reason to select non-Government hospitals was that in such hospitals, the patients and/or their relatives pay directly to hospitals (from their own pocket in most of the cases) for health care services and hence their expectation level is high. The patients and/or relatives evaluate the health care services provided by a hospital in terms of the amount that is paid by them. Usually such patients and their relatives are educated and can raise their voice in social media world.

Online comments were gathered from different social media platforms. The step of gathering of comments was followed by laborious manual task of reading each comment and assignment of class to it. Text mining techniques were applied on the built corpora and the results were analyzed.

The rest of paper is organized as follows:

After the introductory section, literature review is presented in order to introduce reader to academic activities similar and related to the work under discussion. Section III titled "Data Preparation" carries full description of the challenges of construction of corpora that can be used as an input to discover sentiments automatically. Section IV discusses results and intuitive reasoning behind the gained results. The evaluation performance of text mining techniques is also given. Section V concludes this paper along with discussion of future research directions.

II. Literature Review

This section will discuss few of the attempts of research that are made in the field of sentiment analysis till now. After describing these attempts, application of machine learning techniques in the field of health care will be discussed. In this regard, similar works will be also mentioned.

The huge amount of text data available at social media sites provides a great opportunity to individuals as well as different groups. The text data is a mixture of facts and opinions. Even though fake facts and forged opinions exist in the cyber world and it is very difficult to quantify extent of fakeness of internet, the worth of available remaining genuine data cannot be denied. The scope of the paper does not allow the author to discuss this issue further, but several researchers like [1], [2] have made efforts in this area of research. Even after subtracting fake content, things do not become simple. Another issue is fact-opinion mixed content. It is difficult for humans to separate facts from opinions from the text content where opinions are mixed with facts. A biased news available on the print media or social media is an example of fact-opinion mixed content. This issue also does not come in the scope of this paper.

The social media platforms provide opportunity to their users to share their opinions and provide their comments on different issues. The area of webpages where these comments or opinions are written and made available to public, become source of almost-pure opinions (it should be noted that facts are sometimes described in opinion sections) that are precious resources for academia as well as public and private sector. Politicians can find public opinion about different political issues from it. Industry can discover their customers' review in it. Academia can use this data for research purpose. Since the available data is huge, and to deploy human resource to read and summarize these opinions is expensive therefore demand for sentiment-aware applications is great. Nobody from the field expects the machines to be 100% accurate, but even if they are able to produce near-accurate results, it will be enough for decision makers to understand and judge the situation in light of public mood.

Numbers of researchers have conducted researches in order to find the popularity of subject and sentiment analysis, as it is really useful for masses, companies and corporates. Through sentiment analysis, companies can plan for improving themselves and masses can have more insights. If the sentiment analysis is performed taking care of popularity of the subject, then it will be more useful.

In the research world, the notion "sentiment analysis" was firstly used by [3] and other similar term "opinion mining" was first coined by [4]. But the research on the same topic was already started few years ago by [5]–[7].

Document level sentiment analysis was performed by number of researchers including [8], [9]. In the document level sentiment analysis, it is assumed that whole document expresses opinion for single entity. In order to find sentiments at finer level, there was a boom in the field of sentence level sentiment analysis and the main objective of those researches was to find out the sentiment of the each sentence of whole document, which was performed by number of researchers like [10]–[12] . Some work has been done in the field of comparative opinion mining and [13] has done research on YouTube comments for the same purpose. Even in a sentence with single entity, there can be aspect(s) with respect to which there exist positive sentiment(s) and with respect to other aspect(s), the negative sentiment(s) can also exist. Sentence level sentiment analysis provide overall sentiments at sentence level but not at aspect level. Therefore, aspect level sentiment analysis was introduced, which was earlier called as feature level sentiment analysis. The researchers in [14] discussed the usage of state-of-art techniques of CNN and LSTM for the purpose of aspect-level opinion mining. The authors in [15] presented the same issue in context of recommender systems comparison. Many researches on aspect-oriented sentiment analysis were performed via different methods. The investigators in [16] found aspects by frequency of nouns in the whole document and then performed sentiment analysis on retrieved aspects. Author in [17] found relatedness of noun and adjective and via this method tried to retrieve aspect.

III. Data Preparation

Social media platform was the source of data that was input for this research work. The experiences of patients and their relatives regarding health care services, if shared on social media, can be found online at no cost. Through these types of comments, other people can have an idea of hospital's performance. Hospital administration can also use these comments to improve their services, thus enabling them to achieve satisfaction of their patients and attendants. However, there is no straight forward way of achieving this goal because some people post irrelevant comments. For example, they start marketing or branding their products, or they start to post jokes. Such type of comments becomes noise that need to be tackled during preprocessing step. On the other hand, there are relevant comments, which are related to topic. However, the relevant comments come in different varieties. There are various types of relevant comments that were discovered during process of formulation of dataset. Following categories can be constructed after careful study of users' relevant comments:

A. Direct comments with opinion (pointing to topic with opinion)

B. Direct comments without opinion (pointing to topic without any opinion)

C. Informative comments (provide more information on topic)

D. Informative comments with aspects (provide information about aspects of topic)

E. Comparative comments (topic-level comparison)

F. Comparative comments with aspects (aspect-level comparison)

G. Declarative comments.

In this research, the main objective is to **discover** aspect(s) from the comments and then based on extracted aspects, assignment of sentiment scores (positivity, negativity, or objectivity) is needed to be performed. It is possible that in one comment, commentator gives his/her sentiments with respect to multiple aspects. For such type of comments, the need is to extract all the aspects from the comments and then to assign sentiment score based on every extracted aspect. Following two comments are given as an example. The hospital names are replaced by X and Y. It should be noted that the two comments presented as example in following lines, are public comments that are exactly copied here and hence spelling and grammatical mistakes can be found. However, hospital names are replaced. The two example comments are as follows:

1. ***X*** *hospital environment is pretty good, whereas administration are irresponsible*.

2. ***Y*** *nurses are comparatively helpful than* ***X*** *nurses but its parking is very conjusted, particularly for bikers.*

In the first comment, topic or entity is X hospital and aspects are environment and administration. Comment is positive with respect to environment, whereas it is negative in the case of administration. Second comment is not only aspect-oriented but also comparative in nature. The commenter of second comment began with comparison of the nursing service of X hospital and Y hospital and with respect to nursing, comment is positive for Y hospital and negative for X hospital. In the second part of the second comment, entity is "Y hospital" and the aspect in discussion is parking and with respect to "parking" aspect, comment is negative for "Y hospital".

The second example comment is an example of comparative comment with aspects. This research work addresses comments that have/have not aspect(s) based sentiments but inter-hospital comparison does not exist in them.

At the beginning of research, around 10,000 reviews were fetched but all of these reviews/comments were not in textual form. For example, most of the people gave their opinion on the basis of **stars** in comments. For example, 5 stars means "I love this" and 1-star means "I hate this". Such type of reviews was irrelevant for the research. After carefully reading and pruning the comments, less than thousand comments were left that became the subject of the study.

In order to construct dataset consisting of aspect-oriented comments, following tasks were performed:

A. Comments fetching

B. Wiping out noisy and irrelevant comments

In the upcoming sub-sections, discussion about the above tasks will be presented followed by description of process of removing inconsistencies from comments, and after that problem with annotations of comments will be discussed in detail. Experiments, discussions, conclusion and future research will be discussed afterwards.

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Comments Fetching

The first and the foremost step was to gather the comments or reviews of the people. Graph API provided by Facebook, was used to fetch Facebook comments as well as reviews. On the other hand, Twitter API was also used to get tweets of people. However, only few tweets contained discussion of the performance of hospitals and most of the data was fetched via Facebook which was nearly 10000 reviews/comments that was reduced later to less than one thousand comments due to the research work domain constraint. Both providers gave feeds/tweets in various forms. The data was fetched in JSON format and after some processing, was stored in the csv file. To perform these operations, a program was written in JAVA, using which comments were fetched in JSON format and the fetched data was provided to GSON converter (Library written by Google. Inc.), which converted JSON into plain java object (POJO). POJO was read line by line by the program and the data from it was inserted into the CSV file. For the multi-line comments, end of line character was replaced by the space so that every comment fit into single line. The data from hospital's official review page was also fetched.

B. Wiping Out Noisy and Irrelevant Comments

Data which was collected in the first step was not in the usable form and there was irrelevant data also. It is a common practice in Facebook that people tag their friends by typing their names in the comment therefore dataset contained lot of comments and reviews in which names of people were present. Such data was removed manually after reading all comments. Regarding data that was fetched from official hospital's review page, another difficulty and limitation was faced. In those pages, it was not mandatory for people to provide reviews in form of narrations as the field on the form of the webpage was optional for them to fill. The mandatory thing was that they have to give star(s) to provide their feedback, which was not useful for the purpose of this research. Hence such star reviews were also discarded from the input data file. Moreover, there were some comments, which were not relevant. Some people gave marketing comments and some posted jokes. Such comments were also removed manually after reading them. Furthermore, there were some comments, which were written in Roman Urdu language (Urdu language written with the Roman script) to represent opinions in Urdu language. In order to avoid complexity, such comments were also truncated from the CSV file. After above-mentioned preprocessing steps, dataset was ready to be used to perform aspect-based sentiment analysis.

C. Class and Aspect Assignment

For supervised learning, labelled dataset is required that can be used by different classifiers to construct the model that can automatically perform sentiment analysis. In order to label records of the dataset, manual annotation was performed for assignment of class and aspects to the comment/review.

Class assignment was relatively easier than aspect assignment because the whole comment or review was only to be assigned the label of positivity, negativity or objectivity. Whereas, for aspect assignment whole comment was needed to be read to discover aspect(s) after understanding the context and then step of assignment of class (i.e. sentiment) based on discovered aspect, was performed. There were some comments with more than one aspect. In the constructed dataset, maximum of three aspects in a single comment exist. In next lines, few examples are provided to show the complexity of the problem. It should be noted that spelling and grammatical errors can be found in the provided examples. Hospital names in the comments are replaced by symbols X and Y.

1) Class Assignment for Whole Comment

Three classes are usually assigned in normal sentiment classification. From the constructed comments list, example of each class is given below:

a) Positive (Pos): Keep up the good work and pls make better parking arrangements

b) Negative (Neg): Tests is expensive of X Hospital. Normally are twice expensive than any other lab in the city

c) Neutral (Neu): We have to trust doctors even some are bad and some good too.

D. Aspect-based Class Assignment

As discussed earlier, one can provide opinion with respect to more than one aspect in a single comment or review. The aspects are needed to be discovered in the first step followed by the step of sentiment assignment to discovered aspects. Annotation process of assignment of sentiments to aspects recorded class values using following taxonomy:

<Aspect1>_<class>-<Aspect2>_<class>-
<Aspect3>_<class>

Some examples of aspect-based class assignments are given below:

a. Single Aspect Comment Example

In the following example, the commenter has given opinion about quality of healthcare services. The text of the comment is as follows:

I have never ever seen such type of quality healthcare servicess.Simply outstanding...

Here aspect is "services" and assigned class is **positive** hence the label based on used taxonomy will be "services_positive".

b. Double Aspect Comment

In the following example, the commenter has given opinion about performance of hospital with respect to different aspects. The text of the comment is as follows:

I consider **X** hospital to be a hospital full of unprofessional doctors and nurses. You have to micro manage doctors and nurses. Unless you request something (Paging or requesting a doctor) 2-3 times, it won't happen. Dr. **XX** on Special care unit in private section on 3rd floor is the most phathtic and unprofessional doctor I have ever met. He clearly does not like his job. We plan to sue **X** hospital of all the neglect they are doing to our father and I'll make sure that I speak the true colors of **X** in social media in near future to come

Here commenter is complaining about the service of doctor and nurses at X Hospital. For the above comment, the first aspect is doctor and the second aspect is nurse. Hence label will be "doctors_negative-nursing_negative"

Table I presents few aspects that were present in comments of the constructed dataset.

TABLE I. LIST CONTAINING FEW ASPECTS PRESENT IN DATASET

Nursing	Cleaning	Parking
Treatment	Doctors	Care
Environment	Food	Patient
Cafeteria	Facilities	Diagnosis
Management	Expense	Discount

E. Difficulties in Assignments of Sentiments

Annotating sentiments with respect to multiple aspects is marginally difficult than annotating sentiment for entire comment. Various difficulties were faced when annotating sentiments with respect to aspects and entire comment. Some of them are given in following points:

1) It is difficult to understand the polarity of the sentences as well as aspects due to poor usage of English grammar.

2) Too much typos can be present in number of comments.

3) Existence of ambiguity in the comments. For example, the following comment provides the insights to approximate the extent of the problem of ambiguity.

Comment: "Impact of doctor is gt than other and impact of nurse is gt then doctors. It is good for us that sm doctor r gud in **Y** hospital but we cant do ny thng for bad doc".

The above comment is copied from the comments list. It is really difficult even for human being to understand on which side, the polarity of the commenter is. This is the comment which has ambiguity, spelling mistakes and poor usage of English at the same time.

There are two different type of ambiguities in reviews or subjective opinions.

Ambiguity Type – 1:

There were some comments that contained ambiguous statements and it was hard to decide the sentiments of such comments and as a result difficulties were faced while annotating such comments.

Ambiguity Type – 2:

Some comments were written in way that punctuation and grammatical errors and typo mistakes created the impression of presence of ambiguity in them. Hence such comments were apparently ambiguous.

Above example contains both type of ambiguities as it is really hard to decide polarity and there are too many typo and other mistakes in that comment.

F. Finding Aspects

There are four different methods to find the aspect from the text.

1) Extraction based on frequent nouns and noun phrases
2) Extraction by exploiting opinion and target relations
3) Extraction using supervised learning
4) Extraction using topic modeling

The simplest method is method number 1. In this research work, the first method with some modification was used to extract aspects. Custom logic was developed to overcome different problems associated with finding aspects. The algorithm was able to fetch significant number of aspects like doctors, treatment, cafeteria, staff, parking and quality. The algorithm was unable to find few aspects like facilities, nursing, care and management due to low number of comments with such aspects. Algorithm also made some errors in identifying non-aspects as aspects.

G. Sentence Level Sentiment Classification

WEKA was used to perform machine learning task. 10-folds cross-validation was used to test different machine learning algorithm results on the constructed dataset. Number of classes was three namely positive, negative and neutral. Two sentiment lexicons were used to provide sentiment scores of each comment.

H. Aspect-based Sentiment Classification

Two lexicons were used along with different classifiers for aspect-based sentiment classification. 10-folds cross-validation was used in test settings for different experiments. The experiments aimed to find the sentiment of users towards particular aspect of performance of a hospital. Special program was built to extract neighboring words as tokens that were later merged to form new concise comment. Number of classes was two namely positive and negative.

IV. Results and Discussion

A. Sentence Level Sentiment Classification

Experiments were performed with different settings in Weka environment using package for analyzing Affect in tweets[18] and the results of the experiments for sentence level sentiment classification using two lexicons are presented in Table II and Table III. It should be noted that the application of lexicons on the dataset in preprocessing stage resulted in generation of new attributes that carried different scores for comments or tweet. These new attributes were used for classification using different classifiers. Moreover, no tokenization was performed and it was tested that how newly generated attributes help in the sentiment classification process. For example, when lexicon NRC-Affect-Intensity lexicon [19] was applied on the dataset, new attributes that were generated were as follows:

TABLE II. Results of Experiments with Combination of NRC-Affect-Intensity Lexicon and Different Classifiers

Classifier	**Lexicon applied in preprocessing stage**	**Class**	**Accuracy**	**Precision**	**Recall**
Decision Tree J48	NRC-Affect-Intensity lexicon [19] No tokenization	Positive	72.6098 %	0.524	0.418
		Negative		0.784	0.896
		Neutral		0.444	0.200
Naïve Bayes	NRC-Affect-Intensity lexicon [19] No tokenization	Positive	47.2868 %	0.313	0.063
		Negative		0.729	0.541
		Neutral		0.192	0.825
Random Forest	NRC-Affect-Intensity lexicon [19] No tokenization	Positive	76.4858 %	0.587	0.468
		Negative		0.815	0.918
		Neutral		0.591	0.325

TABLE III. RESULTS OF EXPERIMENTS WITH COMBINATION OF SENTIWORDNET LEXICON AND DIFFERENT CLASSIFIERS

Classifier	**Lexicon applied in preprocessing stage**	**Class**	**Accuracy**	**Precision**	**Recall**
Decision Tree J48	SentiWordNet 3.0 [20] No tokenization	Positive	71.8346 %	0.531	0.329
		Negative		0.748	0.940
		Neutral		0.000	0.000
Naïve Bayes	SentiWordNet 3.0[20] No tokenization	Positive	56.8475 %	0.100	0.013
		Negative		0.737	0.731
		Neutral		0.207	0.575
Random Forest	SentiWordNet 3.0[20] No tokenization	Positive	68.9922 %	0.500	0.392
		Negative		0.771	0.866
		Neutral		0.167	0.100

NRC-Affect-Intensity-anger_Score, NRC-Affect-Intensity-fear_Score, NRC-Affect-Intensity-sadness_Score, NRC-Affect-Intensity-joy_Score.

Table II clearly demonstrates that Naïve Bayes is not a suitable classifier to be used when there is no tokenization. Ensemble method of Random Forest outperformed decision tree as expected. Table III shows that SentiWordNet lexicon application on the dataset followed by different classifier usage was not promising as compared to NRC-Affect-Intensity lexicon. Even though decision tree outperformed Random Forest however it can be clearly seen that for Neutral comments, decision tree classification model had no clue for detection of neutral comments. Naïve Bayes performance saw some improvement for the SentiWordNet lexicon as compared to NRC-Affect-Intensity lexicon.

B. Aspect-based Sentiment Classification

After discovery of aspect, the three neighbor words before and three neighbor words after aspect term were taken as the input for the experiment. For example, the comment “X is the best hospital and especially X nursing is the excellent”, has the aspect of “nursing” under discussion. After preprocessing, the processed comment for experiment was “and especially X is the excellent”. The aspect term “nursing” was removed from the comment and three neighbor words before the aspect term and three neighbor words after the aspect term were included for the experiment purpose. For aspect based classification, only positive and negative comments were present in the dataset. Table IV and Table V show the results when the sentiment analysis was applied to discover user sentiments about single aspect. Again no tokenization was performed.

Table IV shows that Random Forest again outperformed other classifiers. Moreover, it can be seen that Naïve Bayes classifier performance has increased as compared to the performance on the full comment. Table V demonstrates an unexpected phenomenon that Naïve Bayes outperformed decision tree and Random Forest classifier. The reason may be the availability of only two scores for the three classifiers and the neighborhood settings for input formulation for experiments of single aspect classification may be more suitable for probabilistic requirements that Naïve Bayes classifier demands. The absence of neutral comments can also be seen as the reason for better performance of Naïve Bayes classifier.

TABLE IV. RESULTS OF EXPERIMENTS FOR ASPECT-BASED SENTIMENT CLASSIFICATION ON COMMENTS DISCUSSING SINGLE ASPECT, WITH COMBINATION OF NRC-AFFECT-INTENSITY LEXICON AND DIFFERENT CLASSIFIERS

Classifier	**Lexicon applied in preprocessing stage**	**Class**	**Accuracy**	**Precision**	**Recall**
Decision Tree J48	NRC-Affect-Intensity lexicon [19] No tokenization	Positive	79.8165%	0.727	0.296
		Negative		0.806	0.963
Naïve Bayes	NRC-Affect-Intensity lexicon [19] No tokenization	Positive	75.2294 %	0.500	0.333
		Negative		0.802	0.890
Random Forest	NRC-Affect-Intensity lexicon [19] No tokenization	Positive	80.7339 %	0.750	0.333
		Negative		0.814	0.963

TABLE V. RESULTS OF EXPERIMENTS FOR ASPECT-BASED SENTIMENT CLASSIFICATION ON COMMENTS DISCUSSING SINGLE ASPECT, WITH COMBINATION OF SENTIWORDNET LEXICON AND DIFFERENT CLASSIFIERS

Classifier	Lexicon applied in preprocessing stage	Class	Accuracy	Precision	Recall
Decision Tree J48	SentiWordNet 3.0 [20] No tokenization	Positive	74.3119 %	0.474	0.333
		Negative		0.800	0.878
Naïve Bayes	SentiWordNet 3.0[20] No tokenization	Positive	77.0642 %	0.563	0.333
		Negative		0.806	0.915
Random Forest	SentiWordNet 3.0[20] No tokenization	Positive	74.3119 %	0.478	0.407
		Negative		0.814	0.854

V. CONCLUSION

Health care services can be evaluated by comments present on social media platforms. Text mining techniques enable automatic discovery of sentiments of opinion givers. This paper described the challenges associated with assessment of performances of hospitals using subjective opinion. It discussed the challenges of formulation and annotation of dataset. It presented how different aspects of health care services can be discovered. It provided results of experiments where sentiment analysis was performed on full comments. Moreover, results were also provided for experiments that aimed to discover sentiment of user for particular aspect of the hospital. In experiments aiming to discover the overall sentiment of the user towards hospital, Random forest and Decision tree classifiers provided good results for the NRC-Affect-Intensity lexicon and SentiWordNet 3.0 lexicons. The experiments that were directed toward finding users' opinion about particular aspect of a hospital, special type of preprocessing was done on input comments and the size of input comments was drastically reduced to maximum of 6 words as a heuristic. The results show that Naïve bayes classifier performance increased drastically reaching to 77.06% using SentiWordNet 3.0 scores. Random forest classifier achieved 80.73% accuracy in the experiments using sentiment scores from NRC-Affect-Intensity lexicon.

In this paper, two sentiment lexicons and three classifiers were used with no tokenization. In future, the work will be enhanced in all directions with inclusion of more lexicons and more classifiers in experiments along with tokenization. In this paper, results of experiment to discover sentiments for single aspect in user comments were presented. In future, the results of experiments that aim to discover sentiment towards multiple aspects of hospital in a single comment will be presented. Depending on the availability of data, one of the prospective area for enhancement of the presented research is the domain of comparative opinion mining where user compares the performance of a hospital with another hospital. Further research in this area is also planned so that machine learning performance in this arena is also explored.

REFERENCES

[1] T. Lavergne, T. Urvoy, and F. Yvon, "Detecting Fake Content with Relative Entropy Scoring.," PAN, vol. 8, pp. 27–31, 2008.

[2] D. M. Lazer et al., "The science of fake news," Science, vol. 359, no. 6380, pp. 1094–1096, 2018.

[3] T. Nasukawa and J. Yi, "Sentiment Analysis: Capturing Favorability Using Natural Language Processing," in Proceedings of the 2Nd International Conference on Knowledge Capture, 2003, pp. 70–77, doi: 10.1145/945645.945658.

[4] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," in Proceedings of the 12th International Conference on World Wide Web, 2003, pp. 519–528, doi: 10.1145/775152.775226.

[5] S. R. Das et al., "Yahoo! for amazon: Sentiment extraction from small talk on the web," in 8th Asia Pacific Finance Association Annual Conference, 2001.

[6] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima, "Mining Product Reputations on the Web," in Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 341–349, doi: 10.1145/775047.775098.

[7] J. Wiebe, "Learning Subjective Adjectives from Corpora," in Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, 2000, pp. 735–740.

[8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up?: Sentiment Classification Using Machine Learning Techniques," in Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, 2002, pp. 79–86, doi: 10.3115/1118693.1118704.

[9] P. D. Turney, "Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews," in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, pp. 417–424, doi: 10.3115/1073083.1073153.

[10] V. Hatzivassiloglou and J. M. Wiebe, "Effects of Adjective Orientation and Gradability on Sentence Subjectivity," in Proceedings of the 18th Conference on Computational Linguistics - Volume 1, 2000, pp. 299–305, doi: 10.3115/990820.990864.

[11] W.-H. Lin, T. Wilson, J. Wiebe, and A. Hauptmann, "Which Side Are You on?: Identifying Perspectives at the Document and Sentence Levels," in Proceedings of the Tenth Conference on Computational Natural Language Learning, 2006, pp. 109–116.

[12] E. Riloff and J. Wiebe, "Learning Extraction Patterns for Subjective Expressions," in Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 2003, pp. 105–112, doi: 10.3115/1119355.1119369.

[13] A. U. R. Khan, M. Khan, and M. B. Khan, "Naïve Multi-label Classification of YouTube Comments Using Comparative Opinion Mining," Procedia Comput. Sci., vol. 82, pp. 57–64, 2016, doi: http://dx.doi.org/10.1016/j.procs.2016.04.009.

[14] W. Quan, Z. Chen, J. Gao, and X. T. Hu, "Comparative Study of CNN and LSTM based Attention Neural Networks for Aspect-Level Opinion Mining," in 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 2141–2150.

[15] M. Hernández-Rubio, I. Cantador, and A. Bellogín, "A comparative analysis of recommender systems based on item aspect opinions extracted from user reviews," User Model. User-Adapt. Interact., pp. 1–61, 2018.

[16] B. Liu, M. Hu, and J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," in Proceedings of the 14th International Conference on World Wide Web, 2005, pp. 342–351, doi: 10.1145/1060745.1060797.

[17] S. Somasundaran, J. Ruppenhofer, and J. Wiebe, “Discourse Level Opinion Relations: An Annotation Study,” in Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, 2008, pp. 129–137.

[18] F. Bravo-Marquez, E. Frank, B. Pfahringer, and S. M. Mohammad, “AffectiveTweets: a Weka Package for Analyzing Affect in Tweets,” J. Mach. Learn. Res., vol. 20, no. 92, pp. 1–6, 2019.

[19] S. M. Mohammad, “Word Affect Intensities,” CoRR, vol. abs/1704.08798, 2017.

[20] S. Baccianella, A. Esuli, and F. Sebastiani, “SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining,” in Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta, 2010.

Priority based Energy Distribution for Off-grid Rural Electrification

Siva Raja Sindiramutty[1], Chong Eng Tan[2], Sei Ping Lau[3]
Faculty of Computer Science and Information Technology
University Malaysia Sarawak, Kota Samarahan, Sarawak

***Abstract*—Rural off-grid electrification is always very challenging due to mostly using limited output renewable energy source such as solar power system. Owing to its nature of power generation that depends on weather condition, the reliability in power provision is often affected by uncontrolled overwhelming usage or bad weather condition. Total power system blackout that frequently happens not only disturb the night activity routine but also can be life threatening if the rural community is unable to initiate telephony communication with the outside world during state of emergency due to power outage. In order to reduce the frequency of total system blackout caused by the reasons mentioned, we proposed a priority-based energy distribution scheme to assist the off-grid standalone solar power system to improve the overall operating hours of the critical appliances in rural areas. The scheme takes into consideration of criticality of the home appliances as defined by the rural users, so that the system would distribute power supply based on the current state of the system with an objective to prolong the service availability of the critical appliances that matter the most to the users. The scheme has been evaluated under simulated scenario and has shown a 100% operation availability of the critical appliance is achievable even during bad weather season that has very low solar input.**

***Keywords*—PI (Panel Input); BP (Battery Power); critical appliances; non-critical appliances; prioritization; operating hour**

I. Introduction

In this modern world, electrical power supply has become so commonly available and it has been a necessity in daily living. Most people living in the urban never have to worry about power provision as there are plenty and cheap. While more energy has been provision to support more applications, the increasing release of greenhouse gases that cause global warming has risen the attention to adopt greener renewable energy sources that have minimal deleterious on nature [1, 2]. On the other hand, while the urban is thinking hard on how to embrace greener renewable energy sources, the rural areas have long been depending on renewable energy power source. The adoption of renewable energy sources such as solar power in rural areas because they are usually not part of the nationwide power grid network hence standalone renewable energy sources are the better option compared to using the highly polluted diesel power generator. The standalone solar power systems have limited power capacity and not very reliable. Many factors such as weather conditions, unpredictable usage load and battery cells deterioration can easily interrupt power availability. In the case of Malaysia especially in the states of Sarawak and Sabah, many of the rural living still rely heavily on standalone power generators powered by diesel fuel but slowly migrating towards renewable energy sources. Rural electrification programs have been initiated by the government to provision off-grid power supply to small villages but owing to rural populations are very scattered and in large quantity, many villages are still waiting for the electrification program to reach them. The wait could be another 5 to 10 years subject to the planning and priority to implement. The standalone off-grid power systems are usually powered through solar diesel hybrid system due to Malaysia located in a region with plenty of sunlight throughout the year and the diesel generators are used as backup during rainy seasons. Renewable energy source such as solar power is an excellent renewable energy source but since it can be easily affected by bad weather or overwhelming usage, there is a need to look into how to strike a balance between the power generated and its usage. Our research works have been investigating the possibility to make optimum use of the amount of power generated in order to improve the overall sustainability of power supply in a more intelligent way. Upgrading and up-scaling the solar power system may solve the power supply issues in short term with an expense of higher cost, but that is not the direction we are looking at. Our approach is to maintain the same solar power system and introduce a new power distribution scheme to prevent total power blackout which happened every time the system running out of power. Our proposed priority-based energy distribution scheme shall ensure critical appliances and emergency communication equipment will still operational at all time without being affected by overwhelming power usage by other non-critical appliances.

This paper has been structured into five sections. Section II explains the background information that leads to the problem that this research is trying to solve. The nature of the solar power system and its importance role in serving the rural areas are also being described. Section III highlights the existing works on power management and schemes by other researchers. Unfortunately, none of the existing works taking rural daily living needs as the main consideration for solution formulation. Section IV describes the design philosophy of the proposed priority-based energy distribution scheme and how it works in ensuring the continuous operation of the critical appliances. Section V presents the simulation results that compare the proposed scheme to standard solar system under different power input and load scenario. The comparison has shown that under simulated environment, total power blackout can be prevented with the adoption of the new scheme.

Section VI concludes the paper by highlighting the impact of the new scheme to rural living.

II. Background

Since solar power is a more convincing energy source in Malaysia owing to its location is closer to the equator and having high yearly irradiance throughout the year, the solar power system has been chosen as the target platform to be enhanced via the new energy prioritize scheme. The greatest advantage of solar power system is that it is renewable daily with sunrise, and it can be implemented almost anywhere within Malaysia [3]. Fig. 1 shows the yearly average irradiance of various locations in Malaysia which indicates relatively high irradiance values that are promising for solar power system implementation.

Solar power system or photovoltaic (PV) system is one of the favorable renewable energy sources in Malaysia. Despite the energy generation efficiency of the system relatively lower than other renewable energy systems, it has the convenience that it is almost pollution free, with relatively small in sustainment and the operating cost is almost free [4, 5]. Each solar cell in a solar panel module generates only about one-half of the electricity and scads of individual solar cells are linked in a sealed, weather proof packages knows as Photovoltaic module [6]. The photovoltaic modules can be connected either in parallel, series or both which known as the Photovoltaic array. This array will eventually connect to power supply module and then to the appliances or to the power grid network. However, solar power systems have some drawbacks in conversion efficiency and the power produced is very much depending on weather condition, scattering of direct sunlight by the atmosphere, tilt angle of the PV solar panels and declination [7, 8]. A standalone off-grid implementation of solar power system for home is illustrated in Fig. 2.

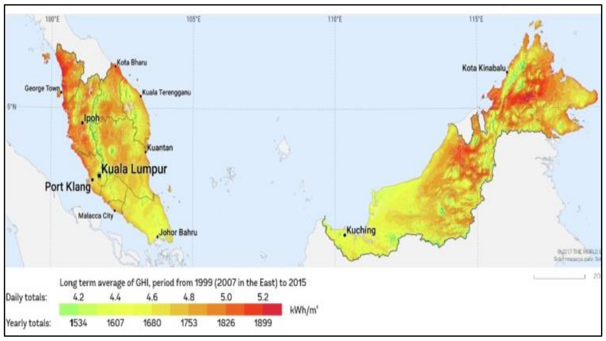

100°E
105°E
110°E
115°E
5°N
Kota Bharu
George Town
Kuala Terengganu
Ipoh
Kuantan
Kuala Lumpur
Port Klang
Malacca City
Johor Bahru
Kota Kinabalu
Kuching
Long term average of GHI, period from 1999 (2007 in the East) to 2015
Daily totals: 4.2 4.4 4.6 4.8 5.0 5.2 kWh/m²
Yearly totals: 1534 1607 1680 1753 1826 1899

Fig 1. Solar Radiation in Malaysia [3].

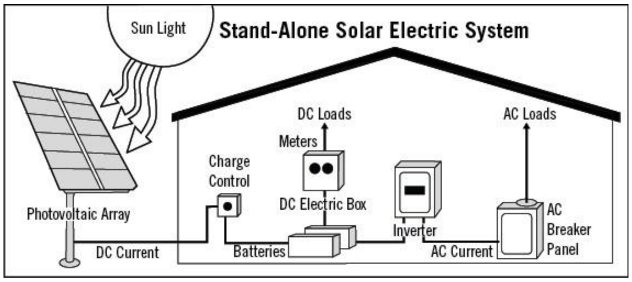


Fig 2. Standalone Solar Power System [9].

The characteristics of such solar power system design have several disadvantages in its application. The power supply is coming from a single source which is from the solar panel to the battery sub-system, where all appliances have equal access to all power available in the system. This is usually the weakest link in the system design that could easily bring down the power system and cause total power blackout. Excessive usage of power by any appliance cannot be controlled by the system and it may use up all reserved power in the battery sub-system at any time. Owing to that, solar power system design will have to take into account the expected loads of the system. An additional of 25% to 30% of excess capacity will also be added to the design to compensate future usage growth. Even so, any sudden increase in usage that exceed the maximum capacity of the system will still trigger power blackout.

This is a commonly problem for power system with a fixed capacity, any unexpected overwhelming usage will just bring down the entire power supply system. This is also the reason many solar system users have the perception that the reliability of solar power system is usually below par. It is not easy to educate users to learn how to self-control in daily power usage, furthermore the manual calculation of conserving power is still too complicated for users with little knowledge in electrical system. Hence, based on the current problem, in order to achieve a better usage experience on a standalone solar power system for home or village, some intelligence can be useful in controlling and regulating the power usage. An intelligently regulated power supply system would be able to prevent frequent system blackout that causes service interruption to all other appliances. Among home appliances, a good power distribution control design shall ensure if not all appliances but some of the more critical appliances such as the basic lighting and emergency communication device to continue to operate even in an event of power shortage.

On top of that, even though solar irradiance is generally high within the region of Malaysia, there are relatively large differences in the amount of yearly irradiance at different locations across the country. The differences in the yearly irradiance would affect the scale and capacity of the solar power system to be implemented. A standardized solar power system design or model may not serve well in all locations with difference yearly irradiance. The frequency of system blackout may be higher in locations with relatively lower irradiance and solar system scale may be over provision in locations with relatively higher irradiance. As shown in Fig. 1 above, the lowest and the highest irradiance differences at different locations can be up to 30%. Also, the solar system reserve capacity design is usually based on the average irradiance which may not accurately capture the actual on-the-day situation. In Table I, we can see the seven days reading of energy collected (watt-hour) at three different locations in Malaysia. The daily differences in daily energy collected can be up to 71% below the weekly average in Kuala Lumpur, 33.5% in Pulau Pinang and 30.8% in Kelantan. With such a significance variation in energy collection between days in the same location, it certainly imposes a great challenge to any standardized solar power system design to cope without experiencing blackout on the specific day of extremely low solar output [10].

TABLE I. SOLAR OUTPUT FOR ONE WEEK IN 3 SELECTED LOCATIONS IN MALAYSIA

	Energy (Wh/Day)		
Day	**Kuala Lumpur**	**Pulau Pinang**	**Kelantan**
1	1145.30	2483.72	3572.67
2	1396.09	2589.24	2809.43
3	2802.60	2404.76	4325.82
4	523.53	3180.18	5907.11
5	1473.49	4326.03	2498.41
6	4115.35	5929.76	3084.14
7	1177.30	4410.37	3081.38

III. RELATED WORK

Throughout the years, intelligent energy management has been the focus of research to maximize and optimize the use of energy for various applications. Saher et al. [11] proposed a priority based maximum consuming power control for smart homes. The proposed system composed of modules such as smart electric sensors (SESs), power provisioning controller (PPC) and home appliances (HA). The PPC functions are to collect ON/OFF information from the connected HA and the immediate power consumption level from the SES, and to send control signal to each HA. After the PPC obtains the data, along with HA priority, it calculates the final target power level of each HA. Those modules are very essential parts of the system, because they control the maximum total power consumption in comprehensive transient behavior considering heterogeneous HAs with different time given. On the other hand, Miroslav & Ales [12] presented design and implementation of a priority based smart home simulation (SHS) system. The proposed system is constructed using the multi-agent approach, where the overall consumption control comes mostly through the inter-agent communication. The presented communication model has been implemented using low performance controllers with limited computational power. Manisa et al. [13] proposed an algorithm for intelligent home energy management (HEM) and to perform demand response analysis for managing high power consumption household appliances. The proposed algorithm handles the home appliances according to their priority and makes sure the total power consumptions are below predefined level. Xin et al. [14] presented a real time household load priority scheduling algorithm based on prediction of renewable source availability. The proposed system is to increase the advantages of renewable energy and reduce the total cost of energy consumption with home users comfort constraints. HA have been allocated dynamic priority based on their different energy consumptions modes and their current status. In every hour, weather condition is taken into account to predict the availability of renewable energy sources. According to the assigned priority, HAs are scheduled based on the predicted output of renewable energy and the forecast electricity. Takekazu et al. [15] comes with a concept of i-Energy as the new energy management algorithm to be aware of efficient and versatile control of e-power flows together with decentralized energy generation and home appliances and offices. The i-Energy idea is best characterized by a new energy management method called Energy on Demand (EoD). Benefits of EoD can achieve the guaranteed minimal energy consumption without jeopardize the quality of living. EoD introduces a new concept which is the explicit demand-based power supply control, a best-effort power distribution method based on appliances priorities and setting ceiling control for power consumption.

Gill et al. [16] invented a ZigBee based home automation system that integrates via the basic home WiFi gateway. The proposed system permits home users to observe and manage the connected appliances in the home, via a plenty of control, including a ZigBee based remote control, and support for WiFi enabled hardware which supports Java. Home users are able to observe and control remotely their home appliances using Internet enable device with Java support. A home gateway is deployed to ease interoperability between heterogeneous networks and gives a consistent interface, regardless of the accessing appliances. Han et al. [17], [18] presented new HEM concept adaptation from ZigBee. Their proposed system in [17] introduces a smart home interfaces and device definitions to permits interoperability together with ZigBee appliances manufactured by different manufactures of electrical appliances, meters and smart energy enabling products. Whereas in [18], a new routing protocol Disjoints Multi Path based Routing (DMPR) is proposed to increase the performance of the ZigBee sensor network. The idea innovates the proposed home energy control systems design that gives intelligent services to home users. Zhao et al. [19] proposed an energy management system for building structures using a multi agent decision making control methodology for building energy management systems (BEMS) for electrical, heating, and cooling energy zones with combined heat and power system optimizations focus at increasing energy efficiency and minimizing the energy cost. Nhat-Hai et al. [20] also initiated the idea for BEMS.

The proposed system is a real time control using wireless sensor network for intelligent BEMS in buildings. Whereas Hiroshi et al. [21] proposed adaptive HEMs/BEMS for controlling energy consumption using the convergence of heterogeneous. Yuvraj et al. [22] presented an occupancy driven energy management for smart building automation that can be used for accurate occupancy detection at the level of individual offices. Using the proposed system, one can achieve potential energy saving of 10% to 15%. Wei et al. [23] proposed a design of energy consumption monitoring and energy saving management system of intelligent building based on the internet of things which has some improvement in the building energy and control, and increases the energy saving of intelligent building. This system is based on wireless network sensors network using the internet of things technology, a detailed analysis of building energy consumption on intelligent building automation systems and appliance, optimum use of good advantages of sensor networks gathers environment data on energy consumption.

Among all works done on intelligence control of power system, none of the design context is focusing on the prioritization of applications for the implementation for rural living and environment. The unique needs of rural living such as basic lighting, difficult in physical access and dependency on remote communication service shall be given a priority access to the very limited resource off-grid power system in order to sustain the service even when the system is under very low energy input.

IV. The Priority-based Energy Distrubution Scheme for Rural Application

A. The Concept

The proposed priority-based energy distribution scheme for rural home appliances have been designed based on the rural living context where several environment factors have been taken into consideration in the overall scheme design. The concept of prioritization is to ensure appliances that are being labeled as critical will continue to operate and survive the total power blackout caused by unpredictable overwhelming usage and also low renewable energy input during rainy seasons. The scheme will be focusing on renewable energy sources, in this case the solar power system, designed for home. Owing to solar power systems are designed based on pre-calculated load profile and usage assumptions, there will be time where the system has to operate outside of these assumptions and calculations. When such time comes, system power blackout is unavoidable as the system design does not give early warning and even if warning is given, how should the users react to it and who should be stopping the use of power in order to prevent unwanted total system blackout. This can be a very complicated situation to deal with among the users on a centralized standalone solar power system. Hence, we proposed a priority-based energy distribution scheme that would decide in advance which appliances will have to go offline earlier than another so that even under the most power constraint situation, the users will still have access to the most critical appliances pre-defined earlier. Under this scheme, the power supply to each home appliance will be categorized into several groups such as critical, semi-critical and standard, or even allocate individual power port for different appliances so that prioritization control can go down to each individual appliance.

Given the required conditions to the central controller of the prioritization controller module, the solar panel input (PI) and the battery capacity (BP) are being monitor closely. The monitored parameters, PI and BP indicate the energy sustainability of the system to support all the connected appliances. Various parameter thresholds are being setup for sustaining different combination of appliances, for example, during the period where the solar is operating under standard designed condition, power will be distributed to all appliances equally, but when the solar input over a period of time is getting critically low, some of the less critical appliances will be cut off from the power supply as a counter measure to conserve energy for other more critical appliances. Under the situation where the battery power is in critically low state, only the most critical appliance is allowed to operate and the rest of the appliances will have to give way to ensure the survival of these devices. In our case here, we define that the telephony is the most critical appliance that should always be kept active for any unforeseen event of emergency. The rural areas have very challenging road assess and lack of communication mean, hence if there is one that is active, it should be kept alive undisturbed. The semi-critical appliance will be the basic lighting for home, which include a very limited number of low power light bulbs to continue lit the house for night activities. The criticality of an appliance is subject to the need of the rural community and their priority can be changed from time to time as deem required.

The priority-based energy distribution scheme will regulate the power so that it will make sure that all the critical appliances will continue to operate by its allocated amount of power. The proposed scheme will have intelligence to separate the critical and non-critical appliances based on their priority so that critical appliances will not be brought down by the high-power consumption usage of non-critical appliances. However, during the period of prolong rainy days, the system may experience great reduction in energy production. In this case, the energy provision will be insufficient for all appliances. Hereby, the proposed scheme shall intelligently cut down power provision for appliances according to its priority for optimum use of available energy. Home appliances can be categorized based on their usage pattern and its importance to the users. For example, the highly demanded appliances will need to operate a longer duration, hence categorizes as higher priority. The appliances that the users cannot live without will have higher priority as well. Owing to that, appliances can be given priority ranking where the appliances with the lowest priority rank will be the first to be cut off from the power system whenever the system starts to run low in its supply. Table II shows an example appliance being categorized with different priority rank as according to the rural community under study.

TABLE II. CRITICAL AND NON-CRITICAL APPLIANCES

Appliances	**Priority**	**Category**
 Telephone	1	Critical appliances
 Network Access	2	
 House Lighting	3	
 TV	4	Non-critical appliances
 DVD Player	5	

Fig. 3 shows the block diagram of the main components of the proposed priority-based energy distribution scheme for rural home appliances. The scope of the rural appliances has been focused on telephone, network access equipment, house lighting, TV and DVD players. The solar power solely depends on the weather condition where the amount of power generated from the solar panels is depending on the sun hours of that particular area and the rainy season has also been taken into consideration for the simulation of charging and discharging of the associated battery sub-system. In principle, the proposed energy distribution scheme is to assist the existing controller system of the conventional solar power system. The scheme is to provide additional intelligence in decision making for the distribution of power to respective appliances. Users' usage pattern refers to the usage of appliances during peak hours where this information will be used for the priority distribution scheme to regulate the appliances based on their criticality, either to allow them to continue operating during peak hours or cut down their power supply during non-peak hours. The power provision is regulated and energy usage is being optimized through the scheme according to the availability of solar power input of the day, as well as the current charge status of battery sub-system.

B. The Priority – based Distribution Scheme

The objective of priority-based energy distribution is to improve the usage effectiveness of the generated renewable energy source. That is, stored energy in the battery sub-system is well used for more critical and meaningful purposes. The distribution scheme is dealing two categories of appliances, critical and non-critical. The critical category contains high priority appliances and the non-critical category contains low priority or common appliances. Fig. 6 shows the flowchart of the priority-based energy distribution scheme and how it prioritizes appliances according to high priority, critical and low priority, non-critical categorization. The scheme will also monitor the power input from solar panels as well as the battery power level in order to priorities the appliances accordingly. At the beginning stage of the scheme, the system will check if either the power input is more than 48 Ampere (A) or the battery power level is in the range between 80 to 100 percent (%), if it does, the system will channel power to all the appliances. This is the comfortable stage of the power system where all appliances regardless of their criticality category, should receive power supply from the system. On the other hand, if either the power input has fall into the range of 41A to 48A or the battery power level are in between 70% to 80%, the system will still channel power to all categories of appliances such as telephone, network access hardware, house lighting and TV. It is to note that the system is now at the lower end of the comfortable stage. Next, if either the power input falls between 26A to 41A or the battery power level goes below 70%, the system now will only channel power to the critical category of appliances where only the telephone, network access hardware and house lighting are operational. Again, if either the power input gets weaker and falls in between 16A and 26A or the battery power level is in between 55% to 60%, the system will start to cut off lower priority appliances in the critical category.

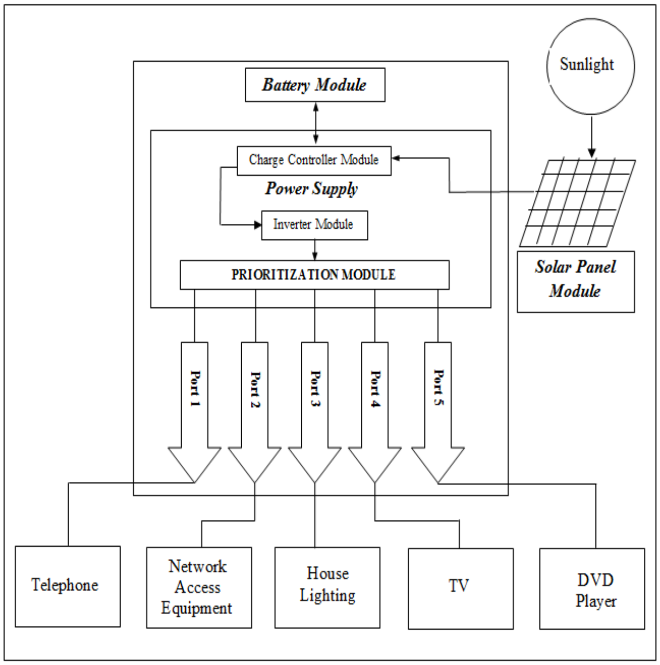


Fig 3. Priority-Based Energy Distribution Scheme for Rural Home.

In this stage, only the highest priority rank appliances, such as the telephone and network access hardware will be operational. Lastly, when either the power input drops below 16A or the battery power level is in between 50% to 55%, the system will only channel power to the highest priority appliance, which is the telephone. If the battery power level hits 50% or lower, the system will enter battery protection cut off stage where the entire power system will stop supplying power completely. This is where the total system blackout will occur as a counter measure for preventing the battery electrode from being damaged so that it will be able to elongate the battery life span. The solar system will now be waiting for energy input from the solar panels in order to bring the battery level beyond 50% so that the system will be back to operational again. This usually happens the next day during sunrise. The most critical appliances such as telephone and network access equipment and house lighting are of essential needs compared to the TV and DVD player, thus the proposed scheme will help so that during event of power shortage, the critical appliances will not be brought down by the usage of less essential non-critical appliances.

V. Simulation Setup, Result and Analysis

A simulation via MATLAB has been carried out to study the impact of the proposed energy prioritization scheme for Smart Rural Home Appliances in terms of their operating hours. The scheme has been simulated to provision power to the respective hardware based on its priority categories where it will react accordingly based on the different condition of a different time of the day as long as it meets the conditions pre-defined. The aim is to improve the operational sustainability of the rural appliances during critical hours as well as increase the effective use of energy generated. The evaluation was made based on the results obtained for comparison to a standalone conventional solar system on sustaining the most needed services and to prevent total power blackout.

A. *Simulation Parameter for Standalone Solar Power and Priority-based Energy Distribution Scheme*

Table III shows a list of the parameters used in the simulation for the comparison between the standalone solar power system and the solar power system with priority-based energy distribution scheme. The performance metrics are based on the total operating hour and their energy efficiency. Both systems have been simulated for a total of 720-unit hours (24 hours x 30 days). The same set of weather condition has been used throughout the simulations. That is a combination of sunny, cloudy and rainy days with weather condition indexes ranging from 0 (lowest) to 8 (highest) represented in the simulations. The power consumption rating value for the telephone, network access equipment, home light, TV and DVD player are adapted from online web simulation carry out by Dicrolic (2011) for energy efficient home appliance in rural area. Power consumption of telephone assumed at 1A, network access equipment is 15A, home lighting is 10A, TV usage is 15A and DVD player is 7A. The performance metrics are measured via the calculation of the average total operating hours of the all appliances, total battery remaining charge (battery charge storage) and the total energy consumed by both standalone solar power system and the solar power system with the proposed scheme.

TABLE III. LIST OF PARAMETERS FOR SIMULATION

Parameter	**Value**
Solar Panel Input (High Intensity)	42 Amp
Weather Condition	(0-8) 0 no sun light – 8 maximum sun light
Diesel Power Generator	60 Amp
Days	30
Hours	24
Battery Capacity (Minimum)	500 Amp Hours
Battery Capacity (Maximum)	1000 Amp Hours
Battery Voltage	12.7 V
Initial Value of Battery Capacity	1000Amp (100 %)
Telephone	1 Amp
Network Access Equipment	15 Amp
Light	10 Amp
TV	15 Amp
DVD	7 Amp
Standard User Usage Pattern Hours	7 p.m. – 12 a.m. per day
Performance Metric Parameters	Appliances Total Operating Hours
	Average Overall Total Battery Charge Storage

VI. RESULT COMPARISONS AND DISCUSSIONS

The output of the randomly generated weather condition and solar panel input in terms of daily and hourly will be shown and discussed in this section followed by the total operating hour for critical and non-critical appliances applied to standalone solar power and solar power system with the proposed energy distribution scheme.

A. *Solar Panels Performance under Different Weather Condition*

Fig. 4 shows the weather condition and solar panel input produced randomly on per hour basis for 30 days (30 set data) for various weather occurrences. The weather condition is an important parameter in this simulation as the performance of the critical and non-critical appliances depends heavily on the weather condition. It becomes a decision maker for the power sources that can be conserved by the solar power system and it also determines how much battery charge storage of the energy prioritization scheme can be charged in order to stay active at all-time especially for the critical equipment. Both parameters were monitored in an hourly basis. The solar panels used in this simulation assumed six solar panel modules rated at 100 watts each. With maximum sun intensity, the output of the solar panel array is 600 watts or 42 amps. As seen in Fig. 4 the shape of the graph of the weather conditions and solar panel inputs are of the same pattern as the solar panel input is dependent on the weather condition in order to generate power which is measured in Ampere per hour. This simulation based on various types of weather condition which is unpredictable by its nature. The graph shows that the weather is uncertain for each day, whereby when the weather graph shows a value of 3.4 it is likely to be a sunny day and if in between the range of 0.6 to 3.4, it is a day with lots of cloud but sunny and finally the range between 0 to 0.5 indicates a rainy day.

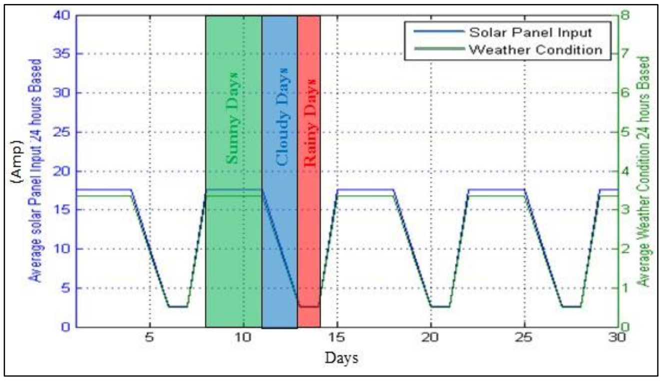


Fig 4. Average Solar Panel Input and Weather Condition.

B. *Prioritization of Rural Appliances under Different Weather Condition*

Table IV shows a test scenario that has been used for evaluating the proposed energy distribution scheme on how it performs prioritized-based energy distribution for all appliances under different weather conditions. The appliances involved in this scenario are telephone, network access, house lighting TV and DVD player, and they have been categorized into either critical or non-critical categories. Both simulated systems receive only one input source which is the solar power. Simulation time for this scenario is from 7 p.m. to 12 a.m. (5 hours). This is the period where all appliances are usually operational. The performance evaluation is determined by the total operating hours of each of the critical and non-critical appliances from 7 p.m. to 12 a.m. The main focus is on the operating hours of the critical category of appliances as its operating hours have great impact to the rural needs. In order to sustain the critical appliances for longer operating hours, the proposed scheme is crucial so that the usage of non-critical appliances will not bring down the critical appliances by causing total power blackout. Thus, the power provision to the critical and non-critical appliances needs to be regulated so that more optimized energy provision and energy conservation can be achieved in the system. Fig. 5 shows the simulation overall results of operating hours for all appliances powered by the two solar power systems for simulation. It can be seen that appliances regulated by the proposed scheme generally have longer operating time compared to the operating time of the standalone conventional solar power system. The operating hours of the conventional solar system also reflects the operating hours of all appliances powered under the same system. Thus, the proposed scheme plays an important role in intelligently provision power from the battery for more critical appliances or services. The simulations have been running in a way that, day 1 to day 4, day 8 to day 11, day 14 to day 17, day 21 to day 24 and day 27 to day 30 are sunny days and all other days are cloudy and rainy.

C. Performance Appliances During Sunny Days and Rainy Days

Fig. 7 compares the operating hours of appliances for the period between 7 p.m. to 12 a.m. on sunny day and rainy day. The operating hours are shown in percentage where 100% means appliance operate fully for 5 hours from 7 p.m. to 12 a.m. Focusing on the critical appliances, the graph showed that, during sunny days, the operating hour percentage for telephone, network access and light can achieve up to 100%. On the other hand, the conventional solar power system only achieved 20% of operating time with all appliances switched on. The proposes scheme has been sustaining all the critical appliances for achieving maximum operating hours by cut downs the power provision for non-critical hours which are TV and DVD player. Thus, the percentage of operating hour for the TV and DVD player from 7 p.m. to 12 a.m. during sunny days is only 55% and 20% respectively. The operating hours of non- critical appliances have been reduced to top up the hours for the critical appliances under the prioritized scheme. On rainy days the battery sub-system cannot be charged to full during day time. Thus, the remaining battery power is low and not able to provision much power for night hours. In During night period, from 7 p.m. to 12 a.m., the battery power level is in critical stage.

The proposed scheme regulated the power provision by channeling more power for the most critical appliance which is the telephone in this case, to maintain its operating hours to 100% whereas all other appliances have been sacrifice to give way to the most critical appliance. Even so, the other lower rank critical appliances such as network access equipment and lighting are still getting 20% of the operating time on a rainy day. On the other hand, it can also be seen that the operating hours of the standalone conventional solar system has operating hours less than 1 hour with all appliances switched on. The solar power system with priority-based energy distribution scheme has demonstrated that it has the ability to sustain operating hours for appliances with critical needs, hence preventing the total power blackout of the system, which usually happened on a non-regulated solar system.

D. Overall Improvement in Appliances Operating Hours

Fig. 8 shows average improvement in terms of operating hours achieved for all appliances regulated by the priority-based energy distribution scheme for the period 7 p.m. to 12 a.m. The average improvement in operating hours for the most critical appliance, the telephone is 462.5%, network access equipment and house lighting achieved average improvement of 293.8%, TV is 125% and DVD player is a reduction of 20.8%. The DVD player has the lowest priority thus the operating hour is expected to be lesser than others relatively more critical appliances. The operating hours for appliances are different under the regulation of the proposed scheme whereas they are the same under the conventional solar power system. In conclusion, the use of intelligent scheme has created a less stressful situation to ensure the most needed appliances to continue to operate without being affected by overwhelming power usage by other appliances. The remaining power in the battery sub-system can be made good use to meet the rural needs under the situation of limited power supply.

TABLE IV. TEST SCENARIO

System	**Priority-based Energy Distribution Scheme**	**Conventional Solar Power System**
Weather Condition	Sunny, Rainy & Cloudy	
Appliances (Critical and Non-Critical)	Telephone Network Access & House Lighting TV and DVD Player	
Input Energy (Only One)	Solar Power (6 Panel)	
Time	7 p.m.-12 a.m. (5 hours)	

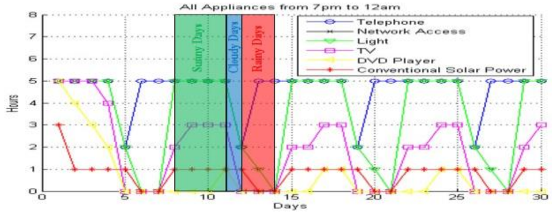


Fig 5. Day to day Performance of Critical and Non-Critical Appliances.

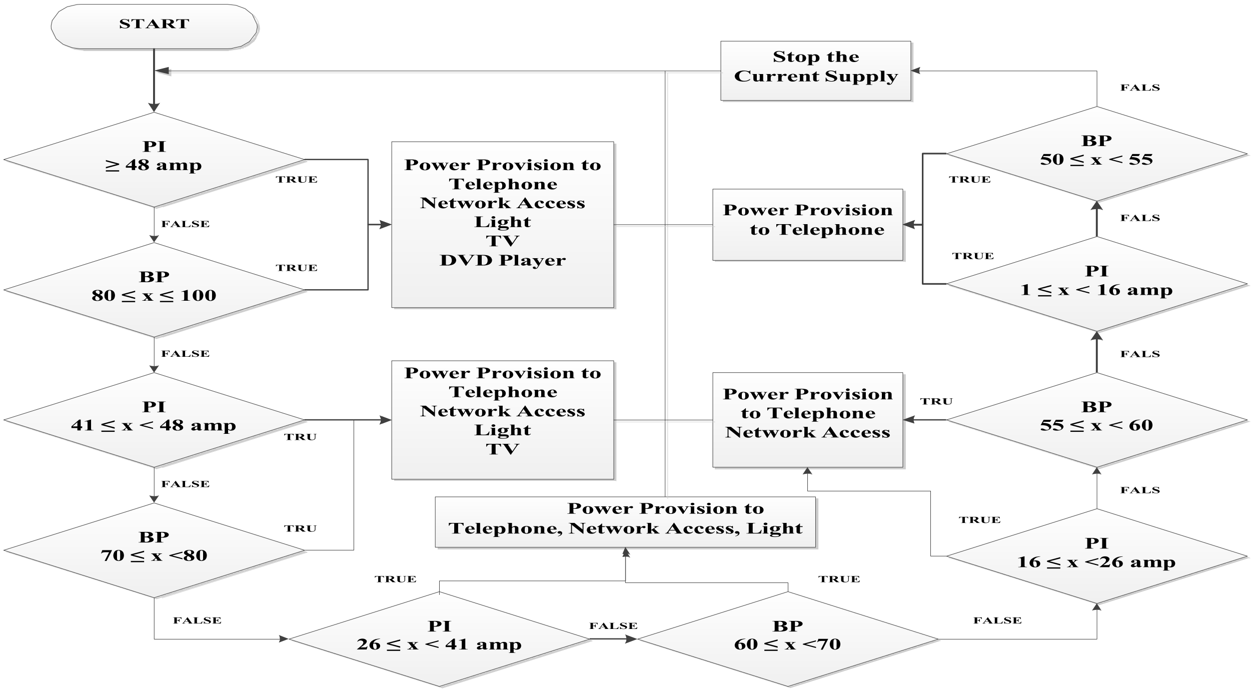


Fig 6. Flowchart of the Priority-Based Energy Distribution Scheme.

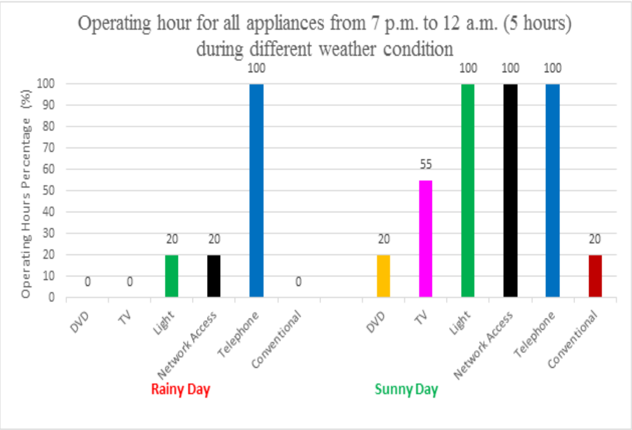


Fig 7. Average Operating Hours of Appliances during Rainy and Dunny Day.

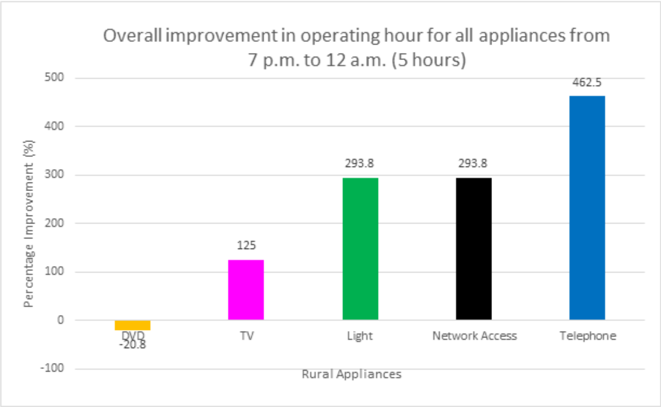


Fig 8. Average Improvement in Operating Hours for Appliances Regulated by the Proposed Scheme.

VII. Conclution

A new energy distribution scheme has been proposed to ease the unreliable off-grid solar power system. The unreliable issue of the solar power system is usually caused by the mismatch between the amount of power generated and the power usage. It has been very difficult to strike a balance for the two without the help of some intelligent scheme. Hence, we have proposed a priority-based energy distribution scheme that take into consideration of critical needs for its formulation. Through the proposed scheme, the issue of overwhelming power usage can be regulated by the system and at the same time conserve enough power for the more critical higher priority appliances to continue to serve the most important needs in rural living. Through the simulated comparison, the impact of the proposed scheme is meaningful as it has created a possible way to prevent total power blackout that will bring down all services either in a home or an entire village. The basic needs of rural living such as lighting and the ability to communicate with the outside world during any life-threatening occasions must be kept alive at all time. From the simulated results, the proposed energy distribution scheme has the ability to prolong the availability of the critical appliances up to 100% availability even under the more severe rainy season scenario. With the new scheme in place, the frequency of total system blackout can be greatly reduced and made the system more reliable.

References

[1] Gaur and G. N. Tiwari, "Exergoeconomic and enviroeconomic analysis of photovoltaic modules of different solar cells," Journal of Solar Energy, DOI: 10.1155/2014/719424, vol. 2014, 8 pages, 2014

[2] P. K. Sahoo, "Exergoeconomic analysis and optimization of a cogeneration system using evolutionary programming," Applied Thermal Engineering, DOI: 10.1016/j.applthermaleng.2007.10.011, vol. 28, no. 13, pp. 1580–1588, 2008.

[3] C. Wei-Nee, “Renewable Energy Status in Malaysia”, Sustainability Energy Development Authority Malaysia, 2012.

[4] Subiyanto, A. Mohammed and M.A. Hannan, “Intelligent photovoltaic maximum power point tracking controller for energy enhancement in renewable energy system,” Journal of Renewable Energy, DOI: 10.1155/2013/901962, vol.2013, Article ID 901962, 9 pages, 2014

[5] T. J. Hammons, J. C. Boyer, S. R. Conners et al., “Renewable energy alternatives for developed countries,” IEEE Transactions on Energy Conversion, DOI: 10.1109/60.900511, vol. 15, no. 4, pp. 481–493, 2000.

[6] O. Gil-Arias and E. I. Ortiz-Rivera, “A general purpose tool for simulating the behavior of PV solar cells, modules and arrays,” in Proceedings of the 11th IEEE Workshop on Control and Modeling for Power Electronics (COMPEL '08), DOI: 10.1109/COMPEL.2008.4634686, pp. 1–5, August 2008.

[7] B. Kamanga, J. S. P. Mlatho, C. Mikeka, and C. Kamunda, “Optimum Tilt Angle for Photovoltaic Solar Panels in Zomba District, Malawi,” Journal of Solar Energy, DOI: 10.1155/2014/132950 vol. 2014, Article ID 132950, 9 pages,2014

[8] E. Calabrò, “An algorithm to determine the optimum tilt angle of a solar panel from global horizontal solar radiation,” Journal of Renewable Energy, DOI: 10.1155/2013/307547 vol. 2013, Article ID 307547, 12 pages, 2013.

[9] U.H. Ibrahim, D.A. Aremu and J.L. Unwaha, “Design Of Stand-Alone Solar Photovoltaic System For Residential Buildings,” in International Journal of Scientific & Technology Research, vol.2, no.2, pp 187-194, 2013.

[10] W.A.Ahmad-Kazwini, “Aplication of Solar Energy in Malaysia,” Faculty of Engineering University of Malaya Kuala Lumpur, 2011

[11] U. Saher, K. Mineo, T. Yasuo and L.O. Azman, “Priority based maximum consuming power control in smart homes,” in IEEE. Innovative Smart Grid Technologies Conference, DOI: 10.1109/ISGT.2014.6816400, pp 1-5, 2014.

[12] P. Miroslav & H. Ales, “Priority-Based smart household power control model,” in IEEE Cons. Electrical Power and Energy Conference, DOI: 10.1109/EPEC.2012.6474976, pp 337- 343, 2012

[13] P. Manisa, K. Muzlu & R. Saifur, “An algorithm for intelligent home energy management and demand response analysis” in IEEE Trans. Smart Grid, DOI: 10.1109/TSG.2012.2201182, vol. 3, no. 4, pp. 2166 – 2173, 2012.

[14] L. Xin, I. Liviu, K. Rui and M. Martin, “Real-time household load priority scheduling algorithm based on prediction of renewable source availability,” in IEEE Trans. Consumer Electronics, DOI: 10.1109/TCE.2012.6227429, vol. 58, no.2, pp. 318-326, 2012.

[15] K. Takekazu, Y.Kenji and M.Takashi, “Energy on Demand Efficient and Versatile Energy Control System for Home Energy Management,” in IEEE Conf. Smart Grid Communications, DOI: 10.1109/SmartGridComm.2011.6102354, pp. 392 – 397, 2011.

[16] K. Gill, S.-H. Yang, F. Yao, and X. Lu, "A zigbee- based home automation system," IEEE Trans. Consumer Electron., DOI: 10.1109/TCE.2009.5174403, vol. 55, no. 2, pp. 422-430, 2009

[17] D.-M. Han and J.-H. Lim, "Smart home energy management system using IEEE 802.15.4 and zigbee," in IEEE Trans. Consumer Electron., DOI: 10.1109/TCE.2010.5606276, vol. 56, no. 3, pp. 1403-1410, 2010.

[18] D.M. Han and J.-H. Lim, "Design and implementation of smart home energy management systems based on zigbee," IEEE Trans. on Consumer Electron., DOI: 10.1109/TCE.2010.5606278, vol. 56, no. 3, pp. 1417-1425, 2010.

[19] P. Zhao, S. Suryanarayanan, & M. G. Simoes, "An Energy Management System for Building Structures Using a Multi-Agent Decision-Making Control Methodology," in IEEE Trans. Industry Applications Society Annual Meeting (IAS), DOI: 10.1109/TIA.2012.2229682, vol. 49. No. 1, pp. 322-330, 2010.

[20] N. Nhat-Hai, T. Quoc-Tuan, J. M. Leger, & V. Tan- Phu, “A real-time control using wireless sensor network for intelligent energy management system in buildings,” in IEEE Workshop. Environmental Energy and Structural Monitoring Systems (EESMS), DOI: 10.1109/EESMS.2010.5634176, pp. 87-92, 2010.

[21] H. Mineno, Y. Kato, K. Obata, H. Kuriyama, K. Abe, N. Ishikawa, & T. Mizuno, “Adaptive Home/Building Energy Management System Using Heterogeneous Sensor/Actuator Networks,” in IEEE Conf. Consumer Communications and Networking Conference (CCNC), DOI: 10.1109/CCNC.2010.5421762, pp. 1-5, 2010.

[22] A. Yuvaraj, B. Bharathan. G. Rajesh, L. Jacob, W. Michael, & W. Thomas, "Occupancy driven energy management for smart building automation," ACM Workshop. On Embedded Sensing Systems for Energy-Efficiency in Building, DOI: 10.1145/1878431.1878433, 2010.

[23] C. Wei & Y. Li, “Design of energy consumption monitoring and energy saving management system of intelligent building based on the Internet of things,” in IEEE Conf. International Conference on Electronics, Communications and Control (ICECC), DOI: 10.1109/ICECC.2011.6066758, pp. 3650-3652, 2011.

Beyond Sentiment Classification: A Novel Approach for Utilizing Social Media Data for Business Intelligence

Ibrahim Said Ahmad[1], Azuraliza Abu Bakar[2], Mohd Ridzwan Yaakub[3], Mohammad Darwich[4]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia, Bangi, Malaysia

Abstract—**Extracting people's opinions from social media has attracted a large number of studies over the years. This is as a result of the growing popularity of social media. People share their sentiments and opinions via these social media platforms. Therefore, extracting and analyzing these sentiments is beneficial in many ways, for example, business intelligence. However, despite a large number of studies on extracting and analyzing social media data, only a fraction of these studies focuses on its practical application. In this study, we focus on the use of product reviews for identifying whether the reviews signify the intention of purchase or not. Therefore, we propose a novel lexicon-based approach for the classification of product reviews into those that signify the intention of purchase and those that do not signify the intention of purchase. We evaluated our proposed approach using a benchmark dataset based on accuracy, precision, and recall. The experimental results obtained prove the efficiency of our proposed approach to purchase intention identification.**

Keywords—Purchase intention; sentiment analysis; lexicon; social media; product reviews

I. Introduction

The internet and web technologies have experienced tremendous development in terms of how data is received, processed and managed over the last decade. The contemporary web provides users with the means to actively interact and modify the contents of the web through social networking platforms. People often share their opinion on these social networking platforms in the form of comments in a blog, debates, and arguments in discussion forums or status updates in social networking channels. The web 2.0 immensely contributed to this development. The web 2.0 provides features that enable users to actively interact and contribute to the web contents rather than merely reading the contents. These features make blogs, Facebook, Twitter and other social networking platforms possible. These platforms that enable people to share their opinion are referred to as social media. Extracting and analyzing the data generated on this social media data is referred to as sentiment analysis [1], [2].

Sentiment analysis research has attracted many studies over the years [3]–[5]. However, most of the research is on the accurate classification of the data into positive, neutral, and negative sentiments. Notwithstanding, several studies have emphasized the varying potentials of sentiment analysis research, in security, tourism, and business intelligence [6], [7]. In this paper, we focus on the potentials of sentiment analysis for business intelligence, specifically on the identification of people's intention to purchase a product, called purchase intention from product reviews.

Consequently, we proposed a lexicon-based approach to classify product reviews whether they signify intention of purchase or not. A lexicon-based approach was selected because it has been applied in other domains with good results. For example a study by [8] focused on cyber-harassment lexicon. In this paper, first we develop a purchase intention lexicon from product reviews, then use the lexicon to classify product reviews as to whether they signify purchase intention or not. We evaluated the approach using benchmark dataset. The experimental results show that it is possible to identify people's purchase intention from product reviews.

The remainder of this paper is organized as follows: Section 2 presents the literature review. Section 3 discusses the proposed method. Section 4 presents the results and discussion while Section 5 presents the conclusion and future works.

II. Literature Review

In the following sub-sections, we discuss the related literature on sentiment analysis and purchase intention identification.

A. Sentiment Analysis

Sentiment Analysis involves the classification of emotions in social media data into positive, negative or neutral sentiment. Sentiment Analysis is possible because of the huge amount of data available through web content, like twitter posts, discussion forums, product reviews, blogs, online markets and comments of web pages [9]–[11]. The task of sentiment analysis is usually achieved through the following steps:

1) Extract the desired content from twitter, blog, and forum.

2) Prepare the extracted data and furnish it by removing irrelevant pieces like symbols and repetition.

3) Detect the sentiment (if any) contained in the contents.

4) Classify the polarity of the contents into positive, negative or neutral.

5) Present the sentiment analysis result.

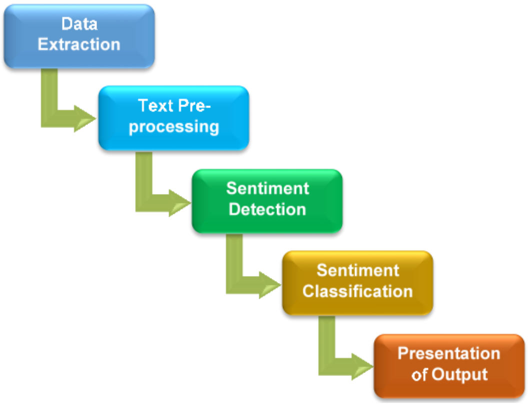


Fig 1. Sentiment Analysis Steps.

Fig. 1 represents the general steps involved in sentiment analysis. Several attempts have been made to achieve the task of sentiment analysis. The main challenge of sentiment analysis is sentiment detection and sentiment classification of the contents [10].

This classification is normally achieved in two main ways, i.e., supervised learning approach and unsupervised learning approach. Supervised learning approach train a sentiment classifier based on the training documents which are represented by the selected features. unsupervised learning approach divide features into three classes, ‘‘positive’’, ‘‘negative’’ and “neutral” based on sentiment lexicon and then count an overall positive/negative score for a document [12].

B. Supervised Learning Approach

Supervised learning approach of sentiment analysis, also known as the machine learning approach of sentiment analysis, involves the use of popular machine learning algorithms for sentiment analysis. This is achieved by training a machine learning algorithm with a labeled dataset and then using that trained algorithm for sentiment analysis [13]. The general methodology for the supervised learning approach is illustrated in Fig. 2 as given by [14]. It involves first creating a training dataset by manually annotating reviews into different sentiment classes, and then use that training set to train a machine learning algorithm so that it can be able to automatically classify new unclassified reviews based on the sentiment they carry.

C. Unsupervised Learning Approach

An unsupervised learning approach of sentiment analysis, also known as the lexicon-based approach involves the use of a list of words with known sentiment value called sentiment lexicon for sentiment analysis. Sentiment lexicon is usually manually developed. The general methodology for the lexicon-based approach of sentiment analysis is illustrated in Fig. 3 as given by [14]. Unsupervised learning approach requires no training set. A review is classified based on developed sentiment lexicon and devised rules.

D. Related Work on Purchase Intention Mining

Purchase intention identification is an important aspect in business intelligence. The Internet and the WWW have provided new avenues through which purchase intention can be identified. Purchase intention online perhaps finds its roots from web usage mining that first appeared in 2000 by [15]. They defined web usage mining as: *“The process of applying data mining techniques to the discovery of usage patterns Web data”*. These patterns are then used for various applications depending on the domain.

In relation to purchase-intention mining specifically, [16] proposed one of the earliest studies. They proposed a Hidden Markov Model (HMM) to predict and internet user's purchase intention based on his online activity data. That is, based on search history, pictures, sounds, and other activities on the web. The precision and recall they got is 51% and 73% respectively. However, with the advent of social media, the data generated on social media has become important for predicting purchase intention. Author in [17] proposed one of the earliest research on the use of social media reviews for purchase intention mining. They proposed a method to automatically identify ‘wishes’ from product reviews by extracting a list of specific list *wish* words from the reviews.

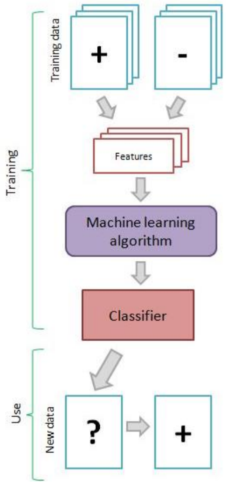


Fig 2. GeneralMethodology for Supervised Learning Sentiment Analysis Approach [14].

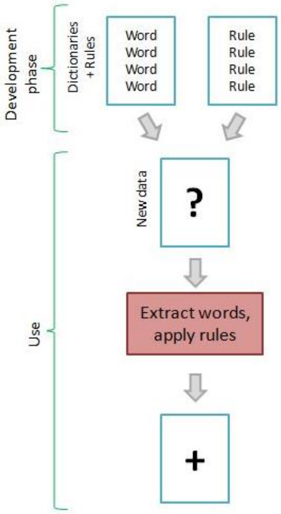


Fig 3. General Methodology for unsupervised Learning Sentiment Analysis Approach [14].

It is important to focus on a specific domain when identifying purchase intention from social media data. This is because different words are used to convey opinions in different domains. Author in [18] proposed a domain-dependent model for identifying user consumption identification from social media data using CNN. They reported an accuracy of 92.54% which is an improvement on previous studies. Author in [19] proposed an approach that uses linguistic features along with statistical features for purchase intention classification. They reported that their proposed approach achieves a significant improvement compared to BOW based features model using Quora post. The best result they obtained is 93% based on AUC. Author in [20] proposed a framework based on the fuzzy set model and association rule mining to predict purchase intention from business companies fan page reviews. They illustrated the effectiveness of their approach using theoretical experiments. Author in [21] proposed an approach based on Recurrent Neural Network (RNN) for purchase intention identification. Their dataset was semi-automatically created from tweets. The RNN model achieved an F-measure of 83% which is better than other models based on linear regression, decision tree random forest, and naive Bayesian algorithms.

III. Proposed Model

A. Dataset

Purchase intention mining from social media is an emerging field of research, therefore sufficient literature in the field is yet to be established. Therefore, there is very little benchmark dataset for purchase intention mining from social media. On product reviews specifically, we are able to find one dataset by [22]. The dataset consists of 7,522 instances, divided as 6,016 for training, 752 for development and 754 for testing. We used this dataset to evaluate our proposed approach.

B. Purchase Intention Classification

In this step, a purchase intention mining approach was proposed. The approach is a classification-based task through which a product review is classified as to whether it signifies purchase intention or not. An unsupervised learning approach was adopted. Therefore, a purchase intention lexicon for movie reviews was developed.

The lexicon was generated using a set of seed words. This was then expanded using therasus.com to include synonyms in a recursive manner exponentially. The synonyms of the synonyms are also included until the list cannot be expanded anymore. The seeds were identified from previous studies. The seed words are *must buy, cannot wait, looking forward, keep an eye on, and must have*. Furthermore, a purchase intention mining approach from product reviews is proposed. The approach uses the lexicon to determine whether a review indicates purchase intention or not. The process is illustrated in Fig. 4.

To determine whether a review indicates purchase intention or not, the developed lexicon is used. A review is classified as indicating purchase intention if it contains any phrase in the seed words, while it does not indicate purchase intention if it does not contain any word in the seed words. The step involved in the classification is given as:

1) Search through a product reviews dataset.

2) For each review, if it contains a phrase from the product review purchase intention lexicon, classify that review as purchase intention review, else, as a review with not purchase intention.

3) Aggregate the total number of reviews that signify purchase intention and reviews that do not signify purchase intention.

C. Evaluation

In sentiment analysis and other classification problems, accuracy, precision, and recall are commonly used in the evaluation of the classification. Therefore, we also used accuracy, precision, and recall in evaluating our proposed approach. In order to understand how these metrics are computed, a knowledge of the confusion matrix is needed. The confusion matrix is a table that shows the performance of a classification task for which the actual values are known. In our case, there are two possible classes, whether a review contains purchase intention or not. Therefore, the 2 * 2 matrix used is shown in Table I.

The number of reviews that are correctly classified as to whether they signify purchase intention or not will be placed in TP and TN respectively, while the number of reviews that are wrongly classified as to whether they signify purchase intention or not will be placed under FP and FN, respectively.

1) Accuracy: Accuracy is a simple evaluation measure calculated as the ratio correctly predicted values to the total values. The equation is given by Equation 1:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

2) Precision: Precision is calculated as the ratio of correctly predicted positive values to the total predicted positive values. Precision tells us how much of the classified data is classified correctly. Precision is given by Equation 2:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

3) Recall: A recall is the ratio of correctly predicted positive values to all values in the actual class. Recall tells us the amount of the correctly classified data; it is given by Equation 3:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

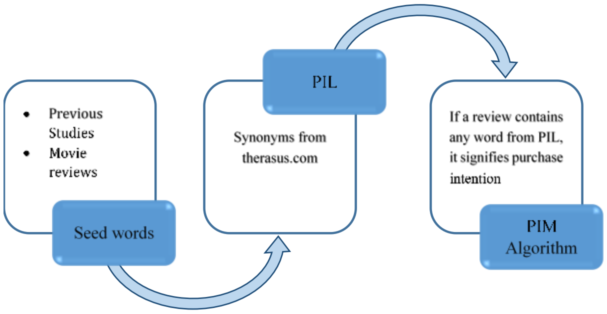

Previous Studies
Movie reviews
Seed words
PIL
Synonyms from therasus.com
If a review contains any word from PIL, it signifies purchase intention
PIM Algorithm

Fig 4. Product Review Purchase Intention Approach.

TABLE I. CONFUSION MATRIX

		Classified Values	
		Positive (PI)	***Negative (Not PI)***
Actual Values	Positive (PI)	True-positive (TP)	False-negative (FP)
	Negative (Not PI)	False-positive (FP)	True-negative (TN)

IV. RESULTS AND DISCUSSION

In this section, we present the experimental results along with a discussion on the results. We used the dataset by [22] to investigate the efficiency of our approach in classifying product reviews based on those that signify purchase intention and those that do not. However, because we are able to get only one benchmark dataset, we randomly divided the dataset into three parts. Then run the experiment on each of the parts, and finally on the entire dataset. This is to be able to study the performance of the proposed approach on different parts of the dataset and deduce a more reliable conclusion. The results of the experiment are presented in Table II.

From Table II, we can see that the proposed is effective in identifying purchase intention from product reviews with an accuracy of 90%, precision of 92%, and a recall of 85%. Similarly, in the three samples of the data created, the accuracy is between 89% and 91%, while the precision is between 88%-100%. However, the recall is low compared to the accuracy and precision, ranging between 75% and 92%. When using the entire dataset, the recall is still the lowest. Recall tells us how many of the reviews that actually signify purchase intention are correctly classified. Therefore, a recall of 85% means that 15% of the reviews that signify purchase intention were not classified correctly. On the other hand, precision tells us the number of reviews classified as signifying purchase intention are actually signifying purchase intention. Therefore, a precision of 92% indicates that only 8% of the reviews that are classified as signifying purchase intention are wrong. Finally, the accuracy tells how accurate our approach is in classifying the reviews are signifying purchase intention or not. An accuracy of 90% means that our approach can classify 90% of the reviews correctly. The results are further illustrated in Fig. 5.

TABLE II. PERFORMANCE OF PROPOSED APPROACH

Data	**Accuracy**	**Precision**	**Recall**
Sample 1	0.89	1	0.75
Sample 2	0.90	0.88	0.88
Sample 3	0.91	0.92	0.92
All	0.90	0.92	0.85

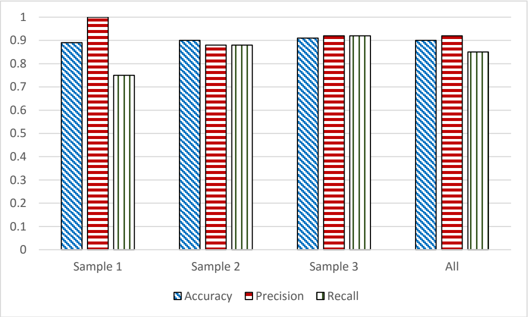


Fig 5. Performance of Proposed Approach.

V. CONCLUSION AND FUTURE WORKS

In this paper, we proposed an approach to automatically classify product reviews into two categories, whether the reviews indicate purchase intention, or whether the reviews do not indicate purchase intention. The approach proposed is a lexicon-based approach, in which a domain-specific purchase intention lexicon was developed and used in the classification. Based on the accuracy, precision, and recall, our approach achieved promising results and hence affirms the notion that reviews contained in social media can be used for business intelligence.

In the future, we intend to use the proposed approach in predicting business performance and the success of a business based on the amount of purchase intention from the reviews.

ACKNOWLEDGMENT

The authors gratefully acknowledge Universiti Kebangsaan Malaysia through grant number DCP-2017-015-1 for supporting this research project.

REFERENCES

[1] J. Awwalu, A. A. Bakar, and M. R. Yaakub, “Hybrid N-gram model using Naïve Bayes for classification of political sentiments on Twitter,” Neural Comput. Appl., pp. 1–14, May 2019, doi: 10.1007/s00521-019-04248-z.

[2] M. R. Yaakub, Y. Li, and J. Zhang, “Integration of Sentiment Analysis into Customer Relational Model: The Importance of Feature Ontology and Synonym,” Procedia Technol., vol. 11, no. Iceei, pp. 495–501, 2014, doi: 10.1016/j.protcy.2013.12.220.

[3] T. Al-Moslmi, N. Omar, S. Abdullah, and M. Albared, “Approaches to Cross-Domain Sentiment Analysis: A Systematic Literature Review,” IEEE Access, vol. 5, pp. 16173–16192, 2017, doi: 10.1109/ACCESS.2017.2690342.

[4] T. Al-Moslmi, M. Albared, A. Al-Shabi, N. Omar, and S. Abdullah, “Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis,” J. Inf. Sci., vol. 44, no. 3, pp. 345–362, Jun. 2018, doi: 10.1177/0165551516683908.

[5] S. R. Ahmad, A. A. Bakar, and M. R. Yaakub, “Metaheuristic algorithms for feature selection in sentiment analysis,” in 2015 Science and Information Conference (SAI), 2015, pp. 222–226, doi: 10.1109/SAI.2015.7237148.

[6] M. Garcia-Herranz, E. Moro, M. Cebrian, N. A. Christakis, and J. H. Fowler, “Using friends as sensors to detect global-scale contagious outbreaks,” PLoS One, vol. 9, no. 4, p. e92413, Apr. 2014, doi: 10.1371/journal.pone.0092413.

[7] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, “Predicting elections with Twitter: What 140 characters reveal about political sentiment,” in Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 2010, pp. 178–185, doi: 10.1074/jbc.M501708200.

[8] M. Rezvan, K. Thirunarayan, S. Shekarpour, V. L. Shalin, L. Balasuriya, and A. Sheth, “A quality type-aware annotated corpus and lexicon for harassment research,” in WebSci 2018 - Proceedings of the 10th ACM Conference on Web Science, 2018, pp. 33–36, doi: 10.1145/3201064.3201103.

[9] B. Liu, Sentiment analysis and opinion mining, vol. 5, no. 1. Morgan & Claypool Publishers, 2012.

[10] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” Ain Shams Eng. J., vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/J.ASEJ.2014.04.011.

[11] D. Zimbra, A. Abbasi, D. Zeng, and H. Chen, “The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation,” ACM Trans. Manag. Inf. Syst., vol. 9, no. 2, 2018, doi: 10.1145/3185045.

[12] S. Wang, D. Li, X. Song, Y. Wei, and H. Li, “A feature selection method based on improved fisher ’ s discriminant ratio for text

sentiment classification," Expert Syst. Appl., vol. 38, no. 7, pp. 8696–8702, 2011, doi: 10.1016/j.eswa.2011.01.077.

[13] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," Expert Syst. Appl., vol. 57, pp. 117–126, Sep. 2016, doi: 10.1016/J.ESWA.2016.03.028.

[14] M. Taboada, "Sentiment analysis: an overview from linguistics," Annu. Rev. Linguist., vol. 2, pp. 325–347, 2016.

[15] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data," ACM SIGKDD Explor. Newsl., vol. 1, no. 2, p. 23, Jan. 2000, doi: 10.1145/846183.846188.

[16] F. Wu, I. H. Chiu, and J. R. Lin, "Prediction of the intention of purchase of the user surfing on the web using hidden Markov model," in 2005 International Conference on Services Systems and Services Management, Proceedings of ICSSSM'05, 2005, vol. 1, pp. 387–390, doi: 10.1109/ICSSSM.2005.1499501.

[17] J. Ramanand, K. Bhavsar, and N. Pedanekar, "Wishful Thinking Finding suggestions and 'buy' wishes from product reviews," in NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text, 2010, pp. 54–61.

[18] X. Ding, T. Liu, J. Duan, and N. J.Y., "Mining user consumption intention from social media using domain adaptive convolutional neural network," in Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[19] V. Gupta, D. Varshney, H. Jhamtani, D. Kedia, and S. Karwa, "Identifying purchase intent from social posts," in Eighth International AAAI Conference on Weblogs and Social Media, 2014.

[20] L.-J. Kao and Yo-Ping Huang, "Predicting purchase intention according to fan page users' sentiment," in 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2017, pp. 831–835.

[21] R. Haque, A. Ramadurai, M. Hasanuzzaman, and A. Way, "Mining Purchase Intent in Twitter," Comput. y Sist., vol. 23, no. 9, 2019, doi: 10.13053/CyS-23-3-3254.

[22] X. Ding, B. Cai, T. Liu, and Q. Shi, "Domain adaptation via tree kernel based maximum mean discrepancy for user consumption intention identification," in IJCAI International Joint Conference on Artificial Intelligence, 2018, vol. 2018-July, pp. 4026–4032.

Efficient Mining of Maximal Bicliques in Graph by Pruning Search Space

Youngtae Kim[1], Dongyul Ra[2]
Computer and Telecommunications Engineering Division
Yonsei University, Wonju, Kangwon, South Korea

***Abstract*—In this paper, we present a new algorithm for mining or enumerating maximal biclique (MB) subgraphs in an undirected general graph. Our algorithm achieves improved theoretical efficiency in time over the best algorithms. For an undirected graph with *n* vertices, *m* edges and *k* maximal bicliques, our algorithm requires O(kn^2) time, which is the state of the art performance. Our main idea is based on a strategy of pruning search space extensively. This strategy is made possible by the approach of storing maximal bicliques immediately after detection and allowing them to be looked up during runtime to make pruning decisions. The space complexity of our algorithm is O(*kn*) because of the space used for storing the MBs. However, a lot of space is saved by using a compact way of storing MBs, which is an advantage of our method. Experiments show that our algorithm outperforms other state of the art methods.**

Keywords—Graph algorithms; maximal bicliques; maximal biclique mining; complete bipartite graphs; pruning search space; social networks; protein networks

I. Introduction

A biclique is a graph (or a subgraph of a graph) whose vertex set can be partitioned into two component sets where every vertex in one set is adjacent to every vertex in the other set. A biclique is also referred to as a complete bipartite graph. A maximal biclique (MB) of a graph G is a biclique which cannot be not a subgraph of another biclique of G.

Nowadays social networks based on the internet or mobile communications are popular [1]. Protein interaction networks receive much attention in biomedical areas [2]. The emerging block chain technology must handle large-scale graphs [3]. In these fields, enumerating all MBs existing (as subgraphs) in a graph is very important to many practical data mining problems. As networks get large in size, efficiency in speed and space of algorithms becomes important.

In this paper, we introduce a new efficient algorithm that can enumerate all MBs in an undirected graph given as input. Henceforth, we use variables *n*, *m* and *k* to denote the number of vertices, edges and MBs in an input graph, respectively. The emphasis in this research is to improve performance of fully general algorithms that involves no constraints. The constraints that can be placed on the algorithms are diverse. Some algorithms accept only bipartite graphs as input. Other algorithms produce only MBs whose component sets are independent sets. There can be size constraints on the component sets. We aim to design a fully general algorithm that does not have any such constraints.

Our approach is based on a new idea of exploiting search space pruning techniques to gain efficiency. In contrast to other fully general algorithms, ours looks up stored MBs to make decisions related to pruning search space, which allows to gain efficiency in time. As a result, we discovered an algorithms with O(kn^2) and O(*kn*) as time and space complexity, respectively.

Our algorithm's time complexity O(kn^2) can be considered to be a significant improvement over the current state of the art O(*kmn*) [4]. The algorithm of Li et al. [4] has been the state of the art for more than a decade and a half among the fully general algorithms. This means that improving speed of the best fully general algorithm has been quite hard. In this respect, contribution of our work is nontrivial.

The theoretical space complexity of our algorithm is O(*kn*) due to the space required to store all MBs. This space requirement seems natural considering the fact that the MBs enumerated anyway need to be loaded into memory to allow application tasks to utilize them. In our scheme of storing MBs, a lot of space can be saved by using a compact way of storing MBs. This is due to the fact that the component vertex sets of different MBs can share their parts and thus the actual amount of space required can be quite less than that of theoretical expectation. This is another advantage of our algorithm. How much space is saved depends on the structure of the graph. It was observed in the experiments that more than 50% of the space is easily saved in case of dense input graphs.

II. Related Work

A lot of research has been done on the problem of mining all MBs in an undirected graph *G*. Algorithms for this purpose belong to one of three categories. Algorithms in the first category have a constraint that the input graph should be bipartite. Algorithms of the other two categories do not have the bipartite-graph constraint. The algorithms in the second category have a restriction that the components of MBs should be independent sets. In other words, they only generate maximal induced bicliques. Algorithms that do not need any constraints or restrictions belong to the third category.

Various algorithms of the first category were developed in the past [5, 6]. Makino and Uno [7] proposed an algorithm whose time complexity is O(n^4) time and O(n^2) space. Zhang et al. [8] recently introduced a novel efficient algorithm of time complexity O(d^2n^2) where *d* is the maximum degree of any vertex. The space complexity is O(min(*d*,*a*)*b*) and *a* and *b* are

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2017R1D1A3B03031855).

cardinalities of the two vertex partitions composing the input bipartite graph *G*.

Many algorithms have also been developed that belong to the second category by allowing a general undirected graph as input. One of them is the algorithm introduced by Dias et al. which requires O(kn^4) time and O(2^n) space [9]. However, this algorithm generates only maximal induced bicliques. If an MB consists of component sets at least one of which is not an independent vertex set, the MB is not enumerated. Kloster et al. [10] pursued improving the algorithm of Dias et al. But their algorithm is specifically designed for general graphs which are near to bipartite graphs. Their algorithm has time complexity of O($knmh^2 3^{h/3}$) where *h* is the cardinality of the vertex set whose deletion from *G* makes *G* a bipartite graph. Sullivan et al. [11] attempted to even further improve the algorithm of Kloster et al. and achieved time complexity of O(*knmh*).

There has been research on developing fully general algorithms belonging to the third category. Liu et al. [12] effectively uses the size constraints on both vertex sets to prune unpromising bicliques and to reduce the search space iteratively during the mining process. The time complexity of the proposed algorithm is O(*kdn*), where *d* is the maximal degree of the vertices. But this algorithm has a size constraint in such a way that only MBs are enumerated whose components' sizes are above a threshold *ms*. One of those fully general algorithms with no constraints was proposed by Alexe et al. [13] which has O(kn^3) and O(*kn*) as time and space complexity, respectively. Another general algorithm in this category is that of Tomita et al. [14] whose time complexity is O($3^{n/3}$). The state of the art algorithm in this category is the one by Li et al. [4] as mentioned in section I. This algorithm has time complexity of O(*kmn*) and space complexity of O(*mn*). However, this space complexity does not include the space for storing the MBs enumerated. When the MBs enumerated need to be stored to be used later by application tasks, their space complexity should be O(*kn*).

III. Preliminary on Maximal Bicliques

We assume that an undirected graph $G = (V, E)$ is given to the algorithm where *V* denotes a vertex set and *E* an edge set. Let *n* and *m* denote the number of vertices and edges in *G*, respectively. We use integers between 1 and *n* to denote vertices. Thus $V = \{1, \dots , n\}$. An edge is represented by a set of two vertices (no order between the two). It is said that a vertex of an edge is adjacent to the other vertex of the edge. We use an adjacency list representation of *G*. In this representation, there is a list $L(v)$ for each vertex *v* in *V* which is an ordered list of vertices which are adjacent to *v*.

Let V_1 and V_2 be disjoint subsets of *V*. If every vertex in V_1 is adjacent to every vertex in V_2, then V_1 and V_2 form a biclique $[V_1, V_2]$ which is a subgraph of *G*. Its vertex set is $V_1 \cup V_2$. Its edge set consists of all edges connecting a vertex in V_1 and a vertex in V_2. We call V_1 and V_2 the component vertex sets. A biclique formed by components V_1 and V_2 becomes a maximal biclique (MB) if there is no vertex set $X \supset V_1$ where *X* and V_2 form a biclique and no vertex set $Y \supset V_2$ where V_1 and *Y* form a biclique. If V_1 and V_2 form an MB, V_2 and V_1 can form an MB. Thus $[V_1, V_2]$ and $[V_2, V_1]$ is actually the same MB. Later in this paper, a specific ordering will be enforced for the two components in writing an MB.

To design our algorithm, we begin with the problem of finding a maximal vertex set which can form a biclique with a given vertex set *X*. This set is called an occurrence set of *X*, which is denoted by Oc(*X*) [4]. Throughout this paper, *V* and *E* denote the vertex and edge set of the input graph *G*, respectively.

Definition 1: An occurrence set of $X \subseteq V$ is $\mathrm{Oc}(X) = \{v \in V \mid v \notin X$ and *v* is adjacent to all vertices in $X\}$.

By the definition of Oc(*X*), it is important to note that Oc(*X*) is a maximal vertex set for given *X*. In other words, there is no vertex set *H* where $H \supset \mathrm{Oc}(X)$ and *H* can form a biclique with *X*.

Theorem 1: Let $X \subseteq V$. $[X, \mathrm{Oc}(X)]$ is a biclique.

Proof: Let us select any vertex $u \in X$ and any vertex $v \in \mathrm{Oc}(X)$. By Definition 1, $(u, v) \in E$. Thus *X* and Oc(*X*) form a biclique. Q.E.D.

Note that $[X, \mathrm{Oc}(X)]$ is a biclique. We need to know whether this biclique is a maximal biclique or not. Closure of a vertex set *X* , Cl(*X*), is a maximal vertex set extended from *X* which forms a biclique with Oc(*X*). It is formally defined in Definition 2. Theorem 2 and 3 provide a method for deciding whether $[X, \mathrm{Oc}(X)]$ is an MB or not.

Definition 2: Closure of a vertex set $X \subseteq V$, Cl(*X*), is the occurrence set of Oc(*X*). I.e., $\mathrm{Cl}(X) = \mathrm{Oc}(\mathrm{Oc}(X))$.

Theorem 2: Let $X \subseteq V$. $[\mathrm{Cl}(X), \mathrm{Oc}(X)]$ is an MB.

Proof: $\mathrm{Cl}(X) = \mathrm{Oc}(\mathrm{Oc}(X))$. Thus Oc(*X*) and Cl(*X*) constitute a biclique by Theorem 1. There is no vertex $v \notin \mathrm{Oc}(\mathrm{Oc}(X))$ which is adjacent to all vertices in Oc(*X*) by the property of an occurrence set.

If we assume that there is a vertex $v \notin \mathrm{Oc}(X)$ which is adjacent to all vertices in Oc(Oc(*X*)), contradiction occurs since v is adjacent to all vertices in *X* and thus it should be in Oc(*X*). Thus Cl(*X*) and Oc(*X*) meet the condition of forming an MB. The theorem holds. Q.E.D.

Theorem 3: Let $X \subseteq V$. Then $X \subseteq \mathrm{Cl}(X)$.

Proof: *X* forms a biclique with Oc(*X*). $\mathrm{Cl}(X) = \mathrm{Oc}(\mathrm{Oc}(X))$ is a maximal set that forms a biclique with Oc(*X*). Thus $X \subseteq \mathrm{Oc}(\mathrm{Oc}(X))$. Q.E.D.

Theorem 4 states that if a vertex is added to a set, the corresponding occurrence set may lose some vertices.

Theorem 4: For any vertex set *X*, and a vertex $v \notin X$, if $Z = \mathrm{Oc}(X \cup \{v\})$, then $Z \subseteq \mathrm{Oc}(X)$.

Proof: *Z* consists of only those vertices in Oc(*X*) which are adjacent to *v*. If there is a vertex *u* in Oc(*X*) which is not adjacent to *v*, *u* does not belong to *Z*. Thus the theorem holds. Q.E.D.

IV. Set Enumeration Tree

A. *Set Enumeration Tree as a Search Space*

A graph with a large number of vertices may have a huge number of MBs. The basic strategy of enumerating all MBs is simple as follows: for each $Y \subseteq V$, compute Oc(Y), Cl(Y) and then enumerate [Y, Oc(Y)] as an MB if Y = Cl(Y). In this strategy, all subsets of V should be tried as Y. Therefore, the search space to find MBs is the power set of V. We use a set enumeration tree as the conceptual model and data structure of the search space [15].

In a set enumeration (SE) tree for a vertex set V, a unique node exists for every subset of V. Each vertex is represented by an integer label. The i^{th} vertex in V is given label i, $1 \le i \le n$. The SE tree for V = {1, ... , 4} is illustrated in Fig. 1. Every node has a vertex label. Every node represents a unique subset of V which is formed by including the vertex labels of all nodes on the path from the root to the node. For example, consider the orange node with label 4 in Fig. 1. This node represents the vertex set {1, 3, 4}. All nodes of the SE tree for V covers all subsets of V. A vertex set and its corresponding node in the SE tree is used interchangeably in this paper.

Note that a node with label b has children with labels from b + 1 to n. For a node X, the set of labels on all child nodes is called its tail set, Tail(X). In Fig. 1, Tail(X) = {2, 3, 4} for the node X = {1}.

A depth-first search (DFS) traversal scheme is used to visit all nodes in an SE tree. Fig 2 illustrates the order of node visits in DFS traversal. When the control arrives at a node for the first time, this is the visit to the node. After the control leaves a node, it may return to the node again later by backtracking from a child node. In this paper, a visit to a node stands for the first visit and not the return caused by backtracking. During the (first) visit to a node, the processing related to the node is performed. This is a kind of preorder traversal. The control at a node moves to the leftmost unvisited child of the node. If a node has no more unvisited child, the control backtracks to the parent of the node.

Definition 3: Relation Prior(X, Y) is true if and only if the visit to node X comes before the visit to node Y during DFS traversal of the SE tree. If Prior(X, Y), it is said that X is prior to Y.

Definition 4: Subtree(Y) denotes the subtree whose root is the node of vertex set Y.

For a given graph G, our algorithm does not explicitly build the SE tree of G. It uses a recursive function to implement the DFS traversal of the implicit SE tree. The basic design of our algorithm is the recursive function Basic_GenMB. Our algorithm is started by invoking the recursive function with an empty set ∅ passed to X.

```
Algorithm Basic_GenMB (X, Oc(X), Tail(X) ):
(1)  if [X, Oc(X)] is an MB, mine it;
(2)  for each v in Tail(X):
(3)     Y ← X ∪ {v} ; Tail(Y) = { u ∈ V | u ∈ Tail(X) and u > v};
(4)     Compute Oc(Y) using Oc(X) and v;
(5)     if ( Oc(Y) ≠ ∅ )
(6)        Basic_GenMB (Y, Oc(Y), Tail(Y) );
```

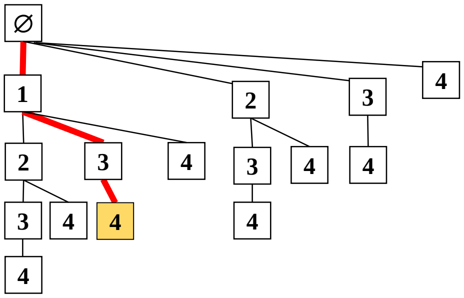


Fig. 1. Set Enumeration Tree with n = 4.

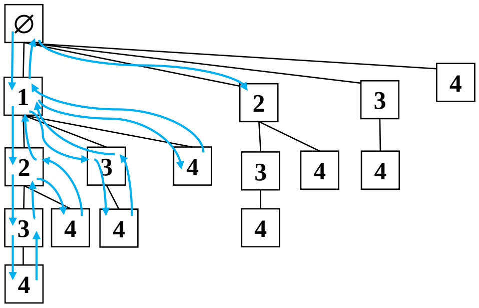


Fig. 2. Depth-First Search Traversal of the SE Tree.

Though significant pruning of search space is done at step 5 of our basic algorithm, more pruning needs to be pursued to improve efficiency. In our algorithm, it is assumed that all vertex sets are ordered sets to improve efficiency in computation. Assume the two vertex sets X and Y form an MB. We write the MB as [X, Y] if Prior(X, Y) is true. Otherwise, we write [Y, X]. In an MB, the component which is prior to the other is the first component and the other the second component. In our algorithm, an MB is mined (i.e. discovered and registered) when its first component is visited. Thus, if [X, Y] was mined before, it means Prior(X, Y) is true and the node X was visited already. When the second component of an MB is visited, the MB is not produced again to avoid duplicate mining.

B. *Storing Maximal Bicliques*

The techniques for achieving efficiency in our algorithm are based upon looking up the MBs already mined during the run of our algorithm. To exploit this idea, it is required to store MBs as soon as they are identified and mined. Our algorithm does not construct the SE tree explicitly. Our algorithm uses an implicit SE tree as the whole search space. When an MB is detected, it should be stored immediately. Its two component vertex sets need to be stored. To store a component, its corresponding node in the SE tree is constructed. The path corresponding to this node is also constructed. The initial part of the path is shared with other existing paths as much as possible.

A tree in which MBs are stored is called an MB tree. The MB tree is a subgraph of the SE tree. The MB tree has only the paths for the components of MBs generated so far. The two nodes representing the components of an MB point to each other by the component pointer (CP). From a node of a component of an MB, the node of the other component can be accessed instantly by using the CP pointer.

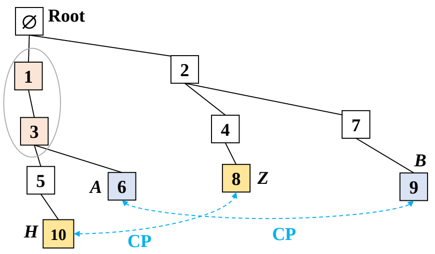


Fig. 3. MB Tree Containing Two MBs: [*H*, *Z*] and [*A*, *B*].

A snapshot of an MB tree is shown in Fig. 3. One MB, [*H*, *Z*], was stored where *H* is {1, 3, 5, 10} and *Z* {2, 4, 8}. Another MB, [*A*, *B*], also exists in the MB tree where *A* = {1, 3, 6}, *B* = {2, 7, 9}. Note that the paths of *H* and *A* share a sub-path consisting of nodes 1 and 3. The paths of an MB tree have the same structure as those of an SE tree. In making decisions about pruning search space, our algorithm needs to look up an MB produced before. Storing and looking up an MB is efficient by adopting the idea of an MB tree.

V. Exploiting Pruning Techniques

We will modify Basic_GenMB to improve efficiency by utilizing MBs stored in the MB tree and pruning search space. Note that MBs are stored as soon as they are identified during the run of the algorithm. An MB is constructed and stored as soon as its first component is visited for the first time during DFS traversal. In this section we will introduce pruning techniques exploited by our algorithm. Our algorithm is a recursive function GenMB. To run our algorithm, the function is invoked as follows: GenMB(∅, *V*, *V*, 0). The algorithm starts at the root of the SE tree. The roles of the parameters are as follows:

- *X*: a vertex set which is being visited by DFS.
- Oc(*X*): the occurrence set of *X*.
- Tail(*X*): the tail set of *X*.
- genflag: If this flag receives 1, an MB should be generated using (extended) *X* and Oc(*X*).

```
Algorithm GenMB ( X, Oc(X), Tail(X), genflag ):
(1)  if (genflag =1 and X ≠ ∅) {
(2)    Closure_extension (X, Oc(X), Tail(X)) ;
(3)    Generate and store MB [X, Oc(X)];
     } // end if
(4)  for each v in Tail(X) do {
(5)      Y ← X ∪ {v} ; Tail(Y) ← {u | u ∈ Tail(X) and v < u} ;
(6)      Compute Oc(Y) using Oc(X) and v;
(7)      if (Oc(Y) = ∅) continue; // Pruning-1
(8)      if (Oc(Y) = 2nd component of MB stored already)
             continue;   //Pruning-2
(9)      if (Oc(Y) = 1st component of MB stored already) {
(10)         Obtain Cl(Y) from node of Oc(Y) using CP link;
(11)         if (Prior(Cl(Y), Y)) continue; // Pruning-3
(12)         else {  Extend Y to Cl(Y) and update Tail(Y);
(13)                 GenMB(Y, Oc(Y), Tail(Y), 0); // Pruning-4
             } // end if
         } // end if
(14)     GenMB(Y, Oc(Y), Tail(Y), 1 ); // Pruning-5
     } // end for
```

If genflag is 1, it means that generation of an MB using *X* and Oc(*X*) is requested. *X* may not be maximal to form an MB. So *X* is extended to its closure Cl(*X*) on line 2. Tail(*X*) is updated by removing vertices added to *X*. More detailed explanation for closure extension and update will be provided later in this section. An MB consisting of Cl(*X*) and Oc(*X*) is generated and stored on line 3. If genflag is 0, it means that an MB composed of *X* and Oc(*X*) was generated previously and thus the MB should not be generated again to avoid duplication. But DFS traversal should continue for nodes in Subtree(*X*). Action "continue" on line 7 and 11 stands for "jumping to next iteration".

The loop from line 4 to 14 is to traverse SE nodes in the subtrees of children of *X*. On line 5, *Y* (a child of *X* in the SE tree) is proposed to be visited next by DFS. *Y* was not visited before. It is visited now. Thus *Y* cannot be a first component of an MB generated earlier. It is necessary to compute Oc(*Y*) (line 6).

A. Pruning scheme 1

Pruning-1: If Oc(*Y*) is an empty set where *Y* is the node to visit next on line 7 of GenMB, Subtree(*Y*) can be pruned.

The first strategy of pruning search space, Pruning-1, is applied on line 7 of GenMB. If Oc(*Y*) is an empty set ∅, there is no need of visiting nodes in the subtree of *Y*. This pruning is possible since the occurrence set of nodes in those subtrees will be ∅ by Theorem 5. The action of "continue" on line 7 makes the algorithm to ignore the remaining part of the loop and start the next iteration (as in C language). This has the effect of ignoring or pruning Subtree(*Y*) during DFS traversal.

Definition 5: If X can be obtained by taking zero or more consecutive elements starting from the first element of an ordered set Z, Prefix(X, Z) is true. Otherwise, Prefix(X, Z) is false.

B. Pruning scheme 2

Pruning-2: If Oc(*Y*) exists in the MB tree as a second component of an MB stored already on line 8 of GenMB, then Subtree(*Y*) can be pruned.

Let us consider a situation to which Pruning-2 can be applied. Fig. 4 has an example. By Theorem 3, Cl(*Y*) is the first component of the MB. *Y* ⊆ Cl(*Y*) by Theorem 2. Cl(*Y*) was visited already before *Y*, which can be derived by the existence of Oc(*Y*) in an MB discovered already. Thus Prior(Cl(*Y*), *Y*). Thus ¬Prefix(*Y*, Cl(*Y*)). Symbol ¬ denotes negation. In this case, Subtree(*Y*) can be pruned safely (on line 8 of GenMB). Theorem 5 proves that this pruning is safe.

Theorem 5: Pruning-2 is safe (This action will not prevent any MB from being generated.)

Proof: Cl(Y) forms a biclique with Oc(Y) by Theorem 3. Y ⊆ Cl(Y). Y ≠ Cl(Y) since Cl(Y) was visited already and Y is being visited now. Thus Y ⊂ Cl(Y). Let v ∈ (Cl(Y) – Y). Note that v is adjacent to all vertices in Oc(Y) since v ∈ Cl(Y).

Y and Oc(Y) form a biclique by definition of an occurrence set. However, Y cannot form a maximal biclique with Oc(Y) since a bigger set (Y $\cup$ {v}) can form a biclique with Oc(Y). This argument can be applied to any node in Subtree(Y).

Consider any node L (other than Y) in Subtree(Y). L is obtained by adding one or more vertices to Y. (For example, consider L in Fig. 4.) L forms a biclique with Oc(L). By Theorem 4, Oc(L) $\subseteq$ Oc(Y). Thus v $\subseteq$ (Cl(Y) – Y) is adjacent to all vertices in Oc(L). Therefore, L$\cup${v} is completely connected with Oc(L). Thus L and Oc(L) can form a biclique but not a maximal biclique because a bigger set $L \cup \{v\}$ can form a biclique with Oc(*L*). Thus no node in Subtree(*Y*) can be a component of a maximal biclique. Q.E.D.

C. Pruning scheme 3

Pruning-3: If Oc(*Y*) exists in the MB tree as a first component of an MB and Prior(Cl(*Y*), *Y*) is true on line 11 of GenMB, then Subtree (*Y*) can be pruned.

Let us consider a situation to which Pruning-3 is applicable. Fig. 5 gives an example of such a situation. Cl(*Y*) exists in the MB tree as a second component of an MB stored already. Line 11 of our algorithm implements this pruning by executing "continue". Theorem 6 below proves that Pruning-3 is safe.

Theorem 6: Pruning-3 is safe (This action will not prevent any MB from being generated.)

Proof: An MB [Oc(Y), Cl(Y)] was stored before. Since Prior(Cl(Y), Y), Cl(Y) was visited before Y. Y $\subseteq$ Cl(Y) by Theorem 3. Y and Cl(Y) cannot be equal since Cl(Y) was visited before and Y is now being visited. So Y $\subset$ Cl(Y).

If Prefix(Y, Cl(Y)) is assumed, Prior(Y, Cl(Y)), which is a contradiction. Thus ¬Prefix(Y, Cl(Y)). There should v $\in$ (Cl(Y) – Y) since Y $\subset$ Cl(Y). (Y $\cup$ {v}) can form a biclique with Oc(Y). So Y and Oc(Y) cannot form an MB.

We can apply the same argument used in the proof of Theorem 5 to conclude that Subtree(Y) can be pruned without missing any MBs. Q.E.D.

D. Pruning scheme 4

Pruning-4: If Oc(*Y*) exists as a first component of an MB stored already and ¬Prior(Cl(*Y*), *Y*) (line 12 of GenMB), then the DFS traversal visits a node *W* in Subtree(*Y*) and all nodes in Subtree(*W*) if and only if Cl(*Y*) $\subseteq$ *W*. This leads to the fact that any Subtree(*W*) contained in Subtree(*Y*) will be pruned if ¬[Cl(*Y*) $\subseteq$ (*W*$\cup$Tail(*W*))].

Fig. 6 shows a situation to which Pruning-4 can be applied. Since Prior(Cl(*Y*),*Y*) is false, *Y* is being visited now but Cl(*Y*) has not been visited yet. But Cl(*Y*) exists as a second component of an MB in the MB tree. Note that *Y* $\subseteq$ Cl(*Y*) by Theorem 3. Thus Prefix(*Y*, Cl(*Y*)). *Y* = Cl(*Y*) or *Y* $\subset$ Cl(*Y*). For example, node *W* in Fig. 6 will be pruned since Cl(*Y*) = {2, 4, 6, 8} is not a subset of *W* = {2, 4, 7} and Tail(*W*) = {8, 9, 10}. Subtree(*W*) will be pruned since any node in it can contain Cl(*Y*). Note that *R* is not visited since *R* = {2, 4, 5}. But, since Tail(*R*) = {6, 7, 8, 9, 10}, some nodes in Subtree(*R*) can contain Cl(*Y*) and thus can be traversed. For example, *Z* = {2, 4, 5, 6, 8} in Subtree(*R*) will be visited since *Z* contains Cl(*Y*). *S* in Subtree(*R*) and all nodes in Subtree(*S*) will be pruned since they cannot contain Cl(*Y*).

Theorem 7: Pruning-4 is safe (This action will not prevent any MB from being generated.)

Proof: Let W be a node in Subtree(Y). Assume that ¬ [Cl(Y) $\subseteq$ W]. Let v be a vertex in Cl(Y) but not in W. Thus v is adjacent to all vertices in Oc(Y). Oc(W) $\subseteq$ Oc(Y) by Theorem 4. Thus v is adjacent to all vertices in Oc(W) because v is adjacent to all vertices in Oc(Y).

Let Q = W $\cup$ {v}. Oc(Q) = Oc(W) because v is adjacent to all vertices in Oc(W). Cl(W) = Oc(Oc(W)) = Oc(Oc(Q)) = Cl(Q). Cl(W) $\supseteq$ Q. Thus Cl(W) $\supset$ W. Therefore, W cannot form a maximal biclique with Oc(W).

Instead, Cl(W) forms an MB with Oc(W). W needs not to be visited. If ¬ [Cl(Y) $\subseteq$ (W$\cup$Tail(W))], then no node Z in Subtree(W) can satisfy Cl(Y) $\subseteq$ Z and thus Subtree(W) can be pruned. Q.E.D.

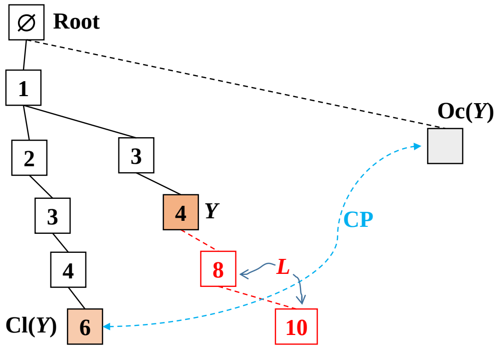


Fig. 4. Situation where Pruning-2 is Applicable.

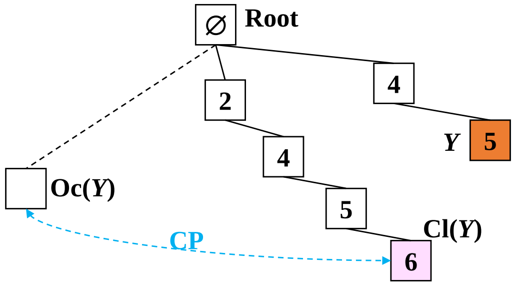


Fig. 5. Situation where Pruning-3 is Applicable.

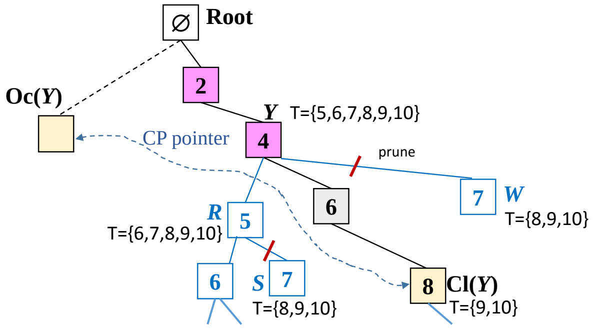


Fig. 6. Pruning-4 is Possible at *Y*; *n* =10, T:Tail.

To implement Pruning-4, our algorithm executes the operations on lines 12, 13 implemented by the next procedure.

```
Procedure for Pruning-4:
(i) // Extend Y to Cl(Y) and update Tail(Y) by using next for loop.
    for each v in Tail(Y):
        if (v in Cl(Y)) { add v to Y; remove v from Tail(Y); }
(ii) GenMB(Y, Oc(Y), Tail(Y), 0 ) ; // recursive invocation of itself
```

This procedure accomplishes pruning search space as stated in Pruning-4. Let us use a simple example in Fig. 7 to understand how our algorithm works related with Pruning-4. Let n = 10. After extension and update operations applied to Fig. 6, $Y = \{2, 4\}$ becomes $Y' = \{2, 4, 6, 8\}$ as in Fig. 7. Tail(Y') = {5, 7, 9, 10}. Note that Tail(Y') has extra vertices 5 and 7 in addition to {9, 10}, the normal tail set of Y'. Thus the tail set becomes unorthodox. GenMB(Y', Oc(Y), Tail(Y'), 0) is called. Zero is passed to genflag to suppress MB generation using Y' and Oc(Y). This call results in visiting only nodes Z in Subtree(Y) where $Z \supset$ Cl(Y) and all nodes in Subtree(Z). The nodes W (in Subtree(Y)) and Subtree(W) will be pruned if $\neg$ [Cl(Y) $\subseteq$ ($W \cup$Tail(W))].

Fig. 8 illustrates nodes in Subtree(Y={2,4}) in the MB tree of Fig. 7 that will be visited by the algorithm. So the parts of Subtree(Y) not covered by traversals in this figure are pruned. First, node Y of Fig. 8(a) and its subtree will be traversed. Secondly, Y of Fig. 8(b) and its subtree will be traversed. Then Y of Fig. 8(c) and its subtree will be traversed. Finally, Y of Fig. 8(d) and 8(e) and their subtree will be traversed.

Theorem 8: Procedure for Pruning-4 accomplishes pruning suggested by Pruning-4.

Proof: Let Y be a node at which Pruning-4 condition is met. We need to verify that Pruning-4 operations achieve the effect that Subtree(W) for any node W in subtree(Y) is pruned if $\neg$ [Cl(Y) $\subseteq$ (W$\cup$Tail(W))].

Extension operation updates Y and Tail(Y) accordingly (see step i). Let Y' and Tail(Y') denote the updated results. Thus Y' = Cl(Y). Tail(Y') contains all elements of Tail(Y) except those added to Y'. Then GenMB(Y', Oc(Y), Tail(Y'), 0) is invoked.

All nodes W that will be visited as a result of this invocation satisfy that Cl(Y) $\subseteq W$ since Y' = Cl(Y) and W is obtained by adding some nodes in Tail(Y') to Y'. Q.E.D.

E. Pruning scheme 5

Pruning-5: If neither Y nor Oc(Y) exist as a component of an MB stored already in the MB tree, the same type of pruning as Pruning-4 should be done. The algorithm will visit a node W in Subtree(Y) and all nodes in Subtree(W) if and only if Cl(Y) $\subseteq W$. As a result, any Subtree(W) contained in Subtree(Y) is pruned if $\neg$[Cl(Y) $\subseteq$ ($W \cup$Tail(W))]

If the conditions required by the pruning strategies introduced so far are not satisfied by Y, our algorithm will arrive at line 14. This happens when neither Y nor Oc(Y) does exist as a component of an MB stored already in the MB tree. (Note that Y cannot be a first component of a stored MB since Y is being visited now and thus was not visited before.) In this case, it is guaranteed that Y is a prefix of Cl(Y), which is proved by Theorem 9. Thus the situation of this Y is quite similar to that of pruning case 4. In both cases, Y is a prefix of Cl(Y). In the current case, Cl(Y) will be a first component while, in case 4, Cl(Y) is a second component of an MB. In the current case, [Cl(Y), Oc(Y)] was not generated and thus will be generated. In case 4, an MB [Oc(Y), Cl(Y)] was generated already. For the current Y, Pruning-5 will be carried out which is similar to Pruning-4.

Note that Oc(Y) was not visited yet. Otherwise, it should exist as a component of an MB in the MB tree. Thus Prior(Y, Oc(Y)) is true. Fig. 9 shows an example of Y on line 14. An MB [Y, Oc(Y)] cannot be proposed as an MB because it is not known yet if Y = Cl(Y) or not. Oc(Y) $\neq \varnothing$ (determined on line 7). Thus Cl(Y) and Oc(Y) form an MB. But the algorithm is visiting node Y. The algorithm invokes a recursive call: GenMB(Y, Oc(Y), Tail(Y), 1). At the start of new instance of GenMB invoked by this call, the two operations are carried out (lines 2, 3 of GenMB): extending Y to Cl(Y) by executing procedure Closure_extension shown below, and generating and storing MB [Cl(Y), Oc(Y)].

Procedure Closure_extension performs the same task of step i of "Proceudre for Pruning-4" shown before. The only difference is that Cl(Y) is not known in this case and thus needs to be computed in using steps (1) and (2). Subtree(W) for any node W in subtree(Y) can be pruned if $\neg$[Cl(Y) $\subseteq$ ($W \cup$Tail(W))] where Y here is before being extended by Closure_extension.

```
Procedure Closure_extension (Y, Oc(Y), Tail(Y)):
(1) Scan all adjacency lists of vertices in Oc(Y);
(2) By scanning, occurrence count is obtained for each vertex;
(3) for each v in Tail(Y):
(4)     if (Count of v = |Oc(Y)|)
            { Add v to Y; Remove v from Tail(Y); }
    end for;
```

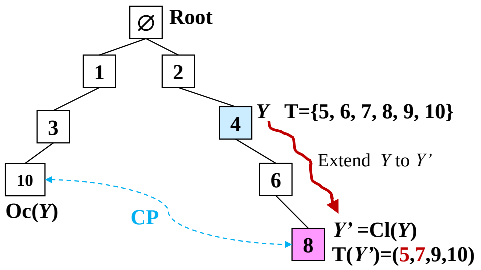


Fig. 7. Extension and update of Y and Tail (Y) to Y' and Tail (Y').

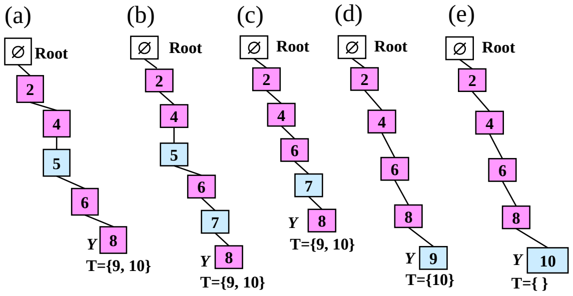


Fig. 8. More Examples of nodes visited while Pruning-4 is Done.

An example of pruning caused by Pruning-5 is shown in Fig. 9 by red line segments. Fig. 10 illustrates additional examples of pruning caused by Pruning-5. Green nodes and their subtrees are traversed; red lines indicate pruning of subtrees. Justifying Pruning-5 can be done using the same arguments used to justify Pruning-4.

Theorem 9: If neither Y nor Oc(Y) exists as a component of an MB stored already, Y is a prefix of Cl(Y).

Proof: A prefix of a node in the SE tree is visited before the node in DFS traversal because of the structure of SE tree. Oc(Y) ≠ ∅ since the test on line 7 was false. Thus Cl(Y) and Oc(Y) can constitute an MB.

If Cl(Y) had been visited already, Oc(Y) should exist as a component of an MB since Cl(*Y*) and Oc(*Y*) can form an MB (by Theorem 2). Because Oc(*Y*) does not exist as a component of an MB in the MB tree, Cl(*Y*) $\supseteq$ *Y* was not visited yet. Therefore, *Y* is a prefix of Cl(*Y*). Q.E.D.

F. *Correctness of our algorithm*

Theorem 10 proves the correctness of our algorithm as an MB enumerator. In other words, there is no MB that is not generated by our algorithm.

Theorem 10: Our algorithm GenMB enumerates all MBs in a graph given as input.

Proof: DFS traversal implemented by GenMB visits all subsets of *V* in an SE tree if there is no pruning. *Y* proposed on line 5 of GenMB is the vertex set being visited.

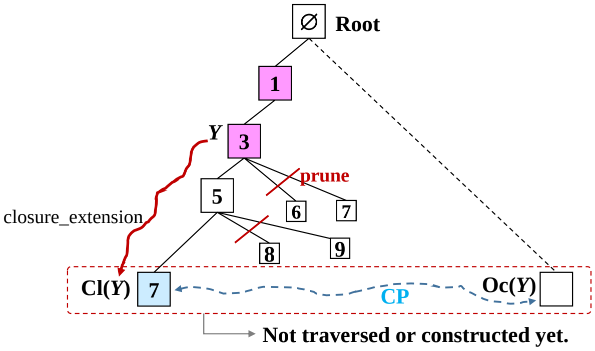


Fig. 9. Situation to which Pruning-5 is Applicable.

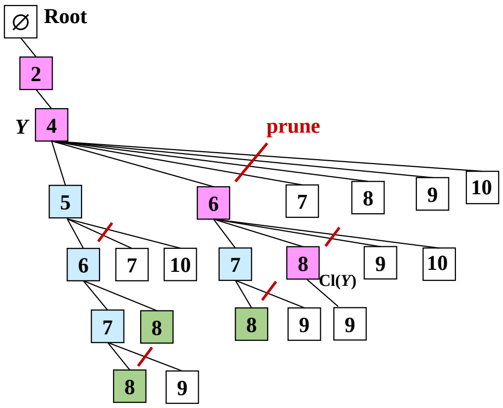


Fig. 10. More Examples for Pruning-5.

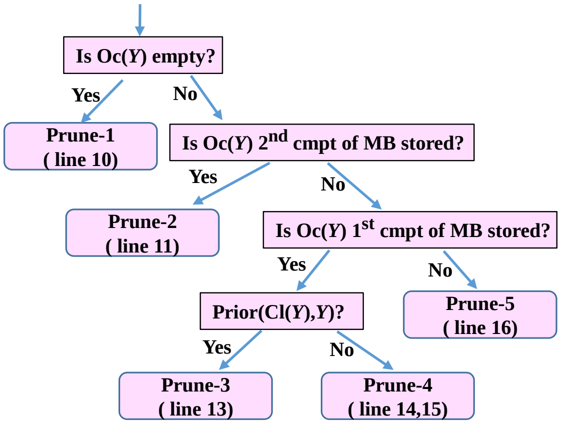


Fig. 11. Decision Tree to Decide Processing for *Y*.

In GenMB, a decision tree shown in Fig. 11 is used to select a case for *Y*. This *Y* will be assigned to one of 5 cases at the leaf in the decision tree. It is necessary to verify that there is no loss of MB after performing actions in any case.

We verified above that the pruning actions for each case are safe and thus no MB is lost from being generated. So the theorem holds. Q.E.D.

VI. Performance Evaluation

In this section, we determine computational complexity of our algorithm GenMB. The result of experimental comparison with current state of the art algorithm is also provided. Let $n = |V|$, and $m = |E|$. Let k be the total number of MBs in the given graph.

A. *Theoretical Performance Evaluation*

Our algorithm uses a recursive function named GenMB. Let $T_{GenMB}(X)$ denote the amount of time required for executing function GenMB when it is invoked with *X* passed to the first input parameter. The time complexity of our algorithm, $T(n, m)$, will be equal to that of $T_{GenMB}(\varnothing)$ which is the time taken by the initial invocation GenMB(∅, *V*, *V*, 0) to finish. An analytical solution for $T_{GenMB}(\varnothing)$ is hard to obtain since systematic progression cannot be formulated for the case of GenMB.

GenMB is invoked in two places in its procedure as follows: (i) line 13 (*X* receives a second component of an MB generated already), and (ii) line 14 (*X* will be a first component of an MB immediately by closure extension.) For each MB, GenMB is called twice (when DFS is visiting its two components). Thus the number of invocations of GenMB is equal to 2*k*. We first compute the time taken by one instance of GenMB, $T_{instance}$, which does not include the time to wait for the return from the recursive call issued during the run of the instance. Then $T(n, m)$ can be obtained from the result of multiplying $T_{instance}$ and 2*k*.

Obtaining $T_{instance}$ requires to compute the time taken by each step of the function. It takes O(*m*) to perform procedure Closure_extension on line 2, which is shown by Theorem 11. Storing an MB involves storing two components in the MB tree and connect them by a CP pointer. By Theorem 12, it takes O(*n*) time to store an MB (on line 3).

To represent a vertex set, say X, a bit-sequence array can be used for fast processing. Let A be the bit-sequence array used to represent X. It has n elements. If $v \in X$, $A[v] =1$. Otherwise, $A[v] = 0$. It takes a constant time to check if a vertex v is in X, since it is needed just to check the value of $A[v]$.

Theorem 11: It takes O(m) time for executing Closure_extension (on line 2).

Proof: Let us consider function Closure_extension. At step 1, adjacency lists of vertices in Oc(Y) are scanned. We prepare an array $Ar[1..n]$ of n integers (initialized to 0) for storing counts of vertices. $Ar[i]$ has the count of vertex i encountered during scanning. During scanning, if an element with value of v is encountered, $Ar[v]$ is incremented. It takes O(m) time for this scanning and counting since the number of elements in adjacency lists is $2m$.

Y and Tail(Y) are represented by bit-sequence arrays. Thus step 4 takes O(1) time. Since $|\text{Tail}(Y)| \leq n$, step 3 and 4 of Closure_extension take $O(n)$ time. Thus time complexity of Closure_extension is O($m + n$). Since m is usually much bigger than n, it can be said that Closure_extension takes O(m) time. Q.E.D.

Theorem 12: It takes O(n) to store an MB in the MB tree.

Proof: Let X be a component of an MB to be stored in the MB tree. Note that the MB tree is a subgraph of the SE tree. Only the nodes in the paths for components of MBs generated are actually constructed. Let $X = \{a_1, a_2, ... , a_r\}$.

Assume that the prefix $(a_1, ... , a_{j-1})$ of the path for X already exists in the MB tree and a_{j-1} does not have a child with label a_j. Let a pointer P point to the root node $\varnothing$ of the MB tree at first. Locate a child of P whose label is a_1. This takes a_1 operations at most since node $\varnothing$ may have children with all labels from 1 to a_1. Let P point to a_1 node.

Among children of a_1, a node with label a_2 should be identified. This may take $(a_2 - a_1)$ operations at most since the node with label a_1 may have children with labels from $a_1 + 1$, ... , a_2. Let P point to a_2 node. By proceeding in this way, P will point to a node of label a_{j-1} eventually. Then searching for node with a_j among children of node with a_{j-1} will be tried but fail, which takes $(a_j - a_{j-1})$ operations at most.

A node with label a_j should be created and attached to node a_{j-1}. For each element in the sequence $(a_{j+1}, ... , a_r)$, a node is created and attached to its predecessor's node. So the number of operations required is $a_1 + (a_2 - a_1) + (a_3 - a_2) + \cdots + (a_j - a_{j-1}) + (r - j) = a_j + r - j$. All a_j, r, j is upper-bounded by n. It takes O(n) to store a component of an MB. Creating a CP link takes O(1). Storing an MB involves storing two components and creating a CP pointer. Thus it takes O(n) to store an MB. Q.E.D.

It takes O(n) time to execute line 5 since scanning Tail(X) can be done in O(n) time. Theorem 13 shows that it takes O(n) to compute Oc(Y) from Oc(X) on line 6.

Theorem 13: Let $Y = X \cup \{v\}$ and $v \in V$. Oc(X) is given. It takes O(n) to compute Oc(Y).

Proof: $\text{Oc(Y)} = \text{Oc(X)} \cap \text{L(v)}$ where the adjacency list $\text{L(v)} = \{u \in V \mid u \text{ is adjacent to } v\}$. Since Oc(X) and L(v) are ordered lists, intersection of them can be done in this way. We use two pointers p1 and p2 to point to elements of Oc(X) and L(v), respectively. Initially they are made to point to the leftmost element of their set. The next loop is repeated until either p1 or p2 falls off the end of their set:

- Use p2 to scan L(v) left to right to find a next element which is not less than the element of p1,
- If the elements of p1 and p2 are the same, the element of p1 is added to the intersection result,
- p1 is made to point to the next element of Oc(X).

Thus computing intersection can be done in |Oc(X)| + |L(v)| steps. Since n is the upper bound of |Oc(X)| and |L(v)|, Oc(Y) can be computed in O(n) time. Q.E.D.

It takes O(1) to carry out the empty-set test on line 7. Theorem 14 shows that it takes O(n) for the look-up operations on lines 8 and 9.

Theorem 14: It takes O(n) time to look up a component of an MB stored already.

Proof: It is needed to find a path corresponding to Y in the MB tree. Let Y = {a1, ... , ar}. It is certain that ar ≤ n. As assumed before, Y is ordered. To locate the nodes for all vertices in Y, the number of nodes to be probed depends only on the final element ar because of the characteristics of the SE tree.

For example, let Y = {2, 5, 8}. The path for Y can be found in this way: we try to find a node with 2 among children of the root, which may require scanning 2 nodes at most (labels 1 and 2); then we try to find a node with label 5 among children of 2 which may require scanning of (5 – 2) nodes at most (with labels 3, 4, 5); we try to find a node with 8 among children of 5, which may require scanning of (8 – 5) nodes at most. In total, the maximum number of operations required is 2 + (5 – 2) + (8 – 5) = 8.

The maximum number of nodes to scan is equal to a_r. Thus the number of operations required is of the order of a_r. Note that $a_r \leq n$. Thus O(n) time is taken at most to locate a node of a vertex set. Finally the node with a_r is checked if it has a CP pointer, which takes O(1) time. Q.E.D.

It takes O(1) time to obtain Cl(Y) on line 10 via CP pointer in the node of Oc(Y). It takes O(n) to test Prior(Cl(Y), Y) since the two lists are scanned using two pointers to find the first position at which the elements do not match. Thus it takes O(n) to execute line 11. Theorem 15 shows that it takes O(n) to do extension and update on line 12.

Theorem 15: It takes O(n) to carry out extension and update operations on line 12.

Proof: The operations to perform is step i of Procedure for Pruning-4. The bit-sequence array representations introduced above are used to implement vertex sets Y, Tail(Y) and Cl(Y). Thus it takes O(1) to test if $v \in$ Cl(Y).

TABLE. I. EXPERIMENTAL RESULTS

Input graphs		**R_1**	**R_2**	**R_3**	**B_1**	**B_2**
graphspec.	n	100	100	300	992	3,890
	m	496	2,476	8,971	1,138	7,729
	k	721	696,000	2.38×10^6	52	5,165
Exec. time (sec)	Li [4]	0.043	269	3,130	1.48	186
	Ours	0.036	99.8	1,474	1.47	65

Adding v to Y takes O(1) since it can be done by setting the corresponding element of the array of Y to 1. Similarly, removing v from Tail(Y) takes O(1). The number of iterations of the loop is not more than |Tail(Y)|. Thus it takes O(n) to execute the loop. Q.E.D.

The amount of time taken to invoke a function on line 13 and 14 requires O(1) time (since the time to wait for the return from the called function is not included).

By adding the amounts of time for lines from 5 to 14, it is found out that one iteration of the loop of line 4 requires O(n) time. This loop iterates |Tail(X)| number of times. |Tail(X)| ≤ n. It takes O(1) time for line 1, O(m+n) for line 2 and O(n) for line 3. Therefore, $T_{\text{instance}} = c_1n + c_2m + c_3n^2$ for some constants c_1, c_2 and c_3. By multiplying 2k and T_{instance}, we obtain time complexity of our algorithm which is T(n) = O(kn^2).

The number of nodes used to store a component of an MB is less than or equal to n. Thus O(n) space is required for storing an MB. So O(kn) space is required to store all MBs. Each active instance of GenMB uses storage for the variables such as X, Oc(X), Y, Oc(Y), Tail(X), Tail(Y), Cl(Y). However, each of them requires O(n) space. The maximum number of instances of GenMB existing in memory at the same time is equal to the height of the SE tree. Thus it is n. So it needs O(n^2) space to store variables used by all active instances of GenMB. Therefore, it requires O(kn+n^2) space by our algorithm. It can be approximated to O(kn) because usually n is much smaller than k.

As a conclusion, we obtain T(n) = O(kn^2) and S(n) = O(kn) as time and space complexity of our algorithm, respectively.

B. Empirical Performance Evaluation

We performed experimentations to confirm the time complexity analysis results of our algorithm. We implemented our algorithm and measured its speed. The current state of the art algorithm is that of Li et al. [4]. This algorithm was implemented to compare its efficiency with ours. We compared the amounts of time required by the two algorithms as shown in Table I. We used five graphs to test the algorithms.

Graphs were generated randomly (marked with R's in Table I) to be used as input. In addition to these artificial graphs, real life protein interaction networks were obtained from the biological repository, BioGrid (marked with B's in Table I) and used as input to test the algorithms [16]. The experimental result shows that our algorithm takes less amount of time than that of Li et al. [4]. This conforms to the theoretical analysis of our algorithm given in the previous subsection. However, if the total number k of MBs is small, efficiency of our algorithm is not manifested fully.

C. Discussions

As far as theoretical time complexity is concerned, our algorithm is superior to any fully general algorithms. The algorithm of Li et al. [4] has been state of the art for more than a decade and a half. Eventually we propose a new state of the art algorithm described in this paper. Another advantage of our algorithm is that a lot of space can be saved by storing all MBs by using paths in a set enumeration tree.

VII. CONCLUSION

In this paper, a new efficient algorithm is proposed for mining all maximal bicliques in an arbitrary undirected graph with n vertices, m edges and k maximal bicliques. The time complexity of ours is O(kn^2) which is a significant improvement over O(kmn) the current state of the art performance [4]. This improvement is made possible by pruning search space extensively in our method. To be able to apply pruning techniques, maximal bicliques are stored as soon as they are discovered. They are looked up to make pruning decisions. Our algorithm requires O(kn) space which is used for storing all MBs. If the MBs need to be loaded into memory after generation to be available for application tasks, any algorithm cannot but require O(kn) space. Because the paths for components of maximal bicliques share a lot of nodes, the actual amount of storage used by our algorithm is less than that expected by theoretical analysis.

Nowadays the networks appearing in the fields of social networks and protein networks have a huge size. Parallelizing the MB-mining algorithms is vital to achieve practical systems [17]. This topic is included in our near future research.

REFERENCES

[1] A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener, "Graph structure in the web," Computer Networks, 33(1-6) pp. 309–320, June 2000.

[2] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, R. Chen, "Topological structure analysis of the protein interaction network in budding yeast," Nucleic Acids Research, vol. 31, no. 9, pp. 2443–2450, May 2003.

[3] C. G. Akcora, M. F. Dixon, Y. R. Gel, M. Kantarcioglu, "Bitcoin risk modeling with blockchain graphs," Economics Letters, vol. 173 pp. 138-142, Dec. 2018.

[4] J. Li, G. Liu, H. Li, L. Wong, "Maximal biclique subgraphs and closed pattern pairs of the adjacency matrix: a one-to-one correspondence and mining algorithms," IEEE Trans. on Knowledge and Data Engineering, vol. 19, no. 12, pp. 1625–1637, Dec. 2007.

[5] M. J. Zaki, C. Hsiao, "Charm: An efficient algorithm for closed itemset mining," In Proceedings of 2nd SIAM International Conference on Data Mining, Arlington, Virginia, pp. 398–416, April 2002.

[6] M. J. Sanderson, A. C. Driskell, R. H. Ree, O. Eulenstein, S. Langley, "Obtaining maximal concatenated phylogenetic data sets from large sequence databases," Molecular Biology Evol., vol. 20, no. 7, pp. 1036–1042, May 2003.

[7] K. Makino, T. Uno, "New algorithms for enumerating all maximal cliques," In Proceedings of 9th Scandinavian Workshop on Algorithm Theory (SWAT 2004), Springer-Verlag, pp. 260-272, July 2004.

[8] Y. Zhang, C. A. Phillips, G. L. Rogers, E. J. Baker, E. J. Chesler, M. A. Langston, "On finding bicliques in bipartite graphs: a novel algorithm and its application to the integration of diverse biological data types," BMC Bioinformatics, vol. 15, no. 110, April 2014.

[9] V.M. Dias, C.M. de Figueiredo, J.L. Szwarcfiter, "Generating bicliques of a graph in lexicographic order," Theoretical Computer Science, vol. 337, pp. 240-248, June 2005.

[10] K. Kloster, A. van der Poel, B. D. Sullivan, " Mining Maximal Induced Bicliques using Odd Cycle Transversals," In Proceedings of the 2019 SIAM International Conference on Data Mining, pp. 324-333, 2019.

[11] B. D. Sullivan, A. van der Poel, T. Woodlief, "Faster biclique mining in near-bipartite graphs," Analysis of Experimental Algorithms, Springer International Publishing, pp 424-453 , Nov. 2019.

[12] G. Liu, K. Sim, J. Li, "Efficient mining of large maximal bicliques," In Proceedings of the 8th international conference on Data Warehousing and Knowledge Discovery, pp. 437-448, Sep. 2006.

[13] G. Alexe, S. Alexe, Y. Crama, S. Foldes, P.L. Hammer, B. Simeone, "Consensus algorithms for the generation of all maximal bicliques," Discrete Applied Mathematics, vol. 145, no. 1, pp. 11-21, Dec. 2004.

[14] E. Tomita, A. Tanaka, H. Takahashi, "The worst-case time complexity for generating all maximal cliques and computational experiments," Theoretical Computer Science, vol. 363 pp. 28–42, Oct. 2006.

[15] R. Rymon, "Search through systematic set enumeration," In Proceedings of 3rd International Conference on Principles of Knowledge Representation and Reasoning, Cambridge, MA, pp. 539-590, Oct. 1992.

[16] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, M. Tyers, "Biological general repository for interaction datasets (BioGRID)," http://thebiogrid.org/download.php.

[17] A. P. Mukherjee, S. Tirthapura, "Enumerating maximal bicliques from a large graph using mapreduce," IEEE Transactions on Services Computing, vol. 10 , no. 5, pp. 771-784 , May 2017.

Nabiha: An Arabic Dialect Chatbot

Dana Al-Ghadhban[1], Nora Al-Twairesh[2]
Information Technology, King Saud University
Riyadh, Kingdom of Saudi Arabia

***Abstract*—Nowadays, we are living in the era of technology and innovation that impact various fields, including sciences. In computing and technology, many outstanding and attractive programs and applications have emerged, including programs that try to mimic the human behavior. A chatbot is an example of the artificial intelligence-based computer programs that try to simulate the human behavior by conducting a conversation and an interaction with the users using natural language. Over the years, various chatbots have been developed for many languages (such as English, Spanish, and French) to serve many fields (such as entertainment, medicine, education, and commerce). Unfortunately, Arabic chatbots are rare. To our knowledge, there is no previous work on developing a chatbot for the Saudi Arabic dialect. In this study, we have developed "Nabiha," a chatbot that can support conversation with Information Technology (IT) students at King Saud University using the Saudi Arabic dialect. Therefore, Nabiha will be the first Saudi chatbot that uses the Saudi dialect. To facilitate access to Nabiha, we have made it available on different platforms: Android, Twitter, and Web. When a student wants to talk with Nabiha, she can download an application, talk with her on Twitter, or visit her website. Nabiha was tested by the students of the IT department, and the results were somewhat satisfactory, considering the difficulty of the Arabic language in general and the Saudi dialect in particular.**

Keywords—Artificial intelligence; natural language processing; chatbot; artificial intelligence markup language; Pandorabots; Arabic; Saudi dialect

I. Introduction

Sciences and technology have a powerful influence on human lives. In recent years, a huge number of achievements have been made in the field of computer science. Artificial intelligence (AI) achieved outstanding and significant results in the field of computing and technology. AI has become particularly important in recent years, and it interacts with many branches of science. AI includes many fields, such as natural language processing (NLP). Natural language refers to a human language (for example, English, Arabic, or Spanish). NLP is a branch of AI that enables human–computer interaction and communication by using a natural language [1]. Several applications have emerged to support NLP. A chatbot is an example of the most interesting artificial intelligence applications that use natural language. A chatbot is a program that enables conducting a human–computer conversation via auditory or textual methods by using a natural language [2]. Hence, it acts as a virtual assistant, and it simulates conversational skills and humanlike behavior through artificial intelligence. In addition, it has an embedded knowledge that facilitates identifying and understanding the sentence and then generates the appropriate response [3]. Nowadays, many chatbots have been developed for various languages and fields. Since 1960, significant work has been done on English chatbots. However, Arabic chatbots are rare owing to the nature and complexity of the Arabic language. In this study, we support the Arabic language by developing a chatbot in the Saudi Arabic dialect. Arabic language is in a state of Diglossia where the formal language used in written form differs radically from the one used in every-day spoken language. The spoken language differs in different Arabic countries producing numerous Arabic dialects i.e. Dialectal Arabic (DA). Until recently, DA was mostly spoken and was never found in written form. The proliferation of social media has changed this trend; as Arabs now use DA in these social media websites. Hence, using DA is more convenient for Arabs; therefore, we propose to use the Saudi dialect for developing Nabiha.

In this study, we propose "Nabiha", a social chatbot that can support conversation with the students of the information technology (IT) department at King Saud University (KSU) using the Saudi Arabic dialect. Nabiha, is available on different platforms: Android, Twitter, and Web. When a student wants to chat with Nabiha, she can either download an Android app, talk with her on Twitter, or visit her website. The aim of Nabiha is to socially interact with the students and answer their enquiries on the courses offered in the IT department or any question related to their academic progress at KSU i.e. it serves as an academic counselor. A usability test was performed to evaluate the Nabiha chatbot.

The remainder of this paper is structured as follows. In Section 2, we describe related work and existing chatbots. Next in Section 3, we present our methodology of chatbot development. Subsequently in Section 4, we describe the implementation of our chatbot. Finally, in Section 5, we show the results of the evaluation of the Nabiha chatbot.

II. Related Work

Chatbots have been developed for many languages and fields. In this section, we describe studies, research, and applications that are similar to our system. First, we review English chatbots in entertainment, medicine, education, and commerce. Additionally, we examine Arabic chatbots and chatbots in other languages.

A. English Chatbots for Entertainment

The primary goal of developing chatbot systems was to mimic and simulate human conversation and amuse users. To achieve this goal, many chatbots have been developed for various purposes. In this section, we present chatbots developed for entertainment. In 1966, the first attempt to build a chatbot was called ELIZA (short for "Elizabeth"), created by

Joseph Weizenbaum at the MIT Artificial Intelligence Laboratory [4]. ELIZA is a chatbot program that enables and facilitates a human–computer conversation in natural language [5]. The idea was simple and based on keyword matching. ELIZA simulated conversation by using "pattern matching." It interacted by using "scripts" written in MAD-Slip (a list-processing computer programming language), which allowed ELIZA to process user inputs and participate in the dialogue by following the rules and instructions from the script. The most famous script used in ELIZA was DOCTOR that simulated a psychotherapist in clinical treatment. The psychiatric interview is an example of a categorized dyadic natural language communication, and this mode of conversation was used [5]. The algorithm of ELIZA is very simple. First, the input is read and inspected for the presence of keywords. When a keyword is found, a sentence is produced according to a rule associated with the keyword. Next, the text is printed. For example, if the user writes a sentence that includes the word "mother," ELIZA's response can be "Tell me more about your family." This response is based on the theory that "mother" and "family" are central to psychological problems, so a therapist should encourage the patient to talk more about their family. Although ELIZA creates responses based on theory, it does not really "understand" this psychological strategy. It only matches the keyword and regurgitates a standard response. In addition, ELIZA has a strategy that is used to keep the conversation going and encourage the patient to reflect and meditate. Specifically, if it cannot find any keyword match, it uses some fixed phrases such as "Very interesting. Please go on." or "Can you think of a special example?" Although ELIZA is an innovative program, it is "memoryless" and does not really "understand" user inputs [6].

Another example is ALICE, a chatbot engine developed by Dr. Richard Wallace in 1995 in SETL language but migrated to Java in 1998 [7]. SETL is an abbreviation for "set language," a high-level programming language with programs coded at an abstract level [8]. Its syntax and semantics are based on the mathematical set theory [9]. ALICE is inspired by ELIZA and its knowledge of English conversation patterns [10]. It uses pattern matching and stores the information in the artificial intelligence markup language (AIML) to form responses to queries [11]. ALICE is available to the public for free under the General Public License (GNU). In 2000, 2001, and 2004 ALICE won the Loebner Prize, an annual competition in artificial intelligence for computer programs that simulate human speech in writing or pronunciation. Although no computer in the contest outperformed humans, she was ranked as "the most human computer" by two panels of judges [6]. In addition, ALICE/AIML engine was ported to several other languages such as C/C++ and PHP, which contributed to the popularity of ALICE and AIML [7]. One of the features that distinguish ALICE is that it provides a powerful capability named symbolic reduction. The purpose of using symbolic reduction is to "jump" from one category to another. For example, two user inputs "Hello, how are you" and "Hi, how are you?" may be matched with two different categories. To map these inputs to a single category ("how are you"), we can use symbolic reduction instead of specifying individual responses for each input [6]. In May 2002, the platform www.pandorabots.com appeared online, allowing Internet users to develop and host their own chatbots based on AIML for free. It currently hosts more than 206,000 chatbots [7]. We use this platform to develop the Nabiha chatbot.

B. Medical English Chatbots

An example of a chatbot in the medical field is Pharmabot developed in 2015 [12]. The Pharmabot is a pediatric medical chatbot that plays a role of a consultant pharmacist. It conducts conversations with users to collect information. Based on the collected information, Pharmabot suggests appropriate and safe generic medications for children. It was developed by using Visual C# as its front-end and Microsoft Access as its back-end, and it was intended to run on a stand-alone computer. The researchers used the left and right parsing algorithm (bottom-up and left-right approach) in their study to reach the desired result. The Pharmabot's main menu includes four buttons: "Start," "Instruction," "Guidelines," and "Exit." When a user clicks on "Instruction," the system displays the procedures on how to access the program. If the user clicks on the "Guidelines" button, the system displays the rules of input/question format. When the user intends to start a conversation with Pharmabot, he/she will click on the "Start" button. Moreover, a dictionary database in Pharmabot explains technical and medical terms for a novice user.

The chatbot was evaluated by two groups of respondents: eight expert pediatricians from St. Vincent Hospital and twenty-eight pharmacy students from Lady Fatima University in the Philippines. Pharmabot's efficiency was measured by a questionnaire with four sections: user-friendliness, appropriateness, consistency, and speed of response. In the survey, the pharmacy students concluded that user-friendliness and consistency of responses of Pharmabot were acceptable ("strongly agree" for both criteria), as were appropriateness of answers and speed of responses ("agree" for both criteria). The experts concluded that Pharmabot was acceptable in terms of user-friendliness, appropriateness of answers, speed of responses, and consistency of responses ("agree" for all criteria). Hence, Pharmabot may be helpful for parents of children who need assistance in choosing the right medication.

Recently, in June and July of 2019, two medical chatbots were developed. The first one was a virtual medical assistant[13], which is a web application with a chatbot module in it. There are two types of users—patient and admin. The knowledge of this chatbot was stored in a SQLite database and used keyword matching patterns to fetch responses to the patient's query. The application allows the patient to log in and enter their symptoms manually, then, the application will predict the diseases through keyword matching patterns. In addition, the chatbot will store the conversation to help the admin add new patterns to the knowledge base. Also, the application can display the nearby physicians to consult, and allow patients to book an appointment with a physician. On the other hand, in July of 2019, a Diabot chatbot "DIAgnostic catboat" was developed with two models, one for the diagnosis of a generic disease, and the second for a more specific diabetes prediction model [14]. Also, they used the static general health and Pima Indian diabetes datasets. The new contribution in this study is using a meta-classifier, which combines a myriad of weaker models and averages them to produce one final balanced and accurate model. In addition, they used React UI and RASA NLU. They concluded that the larger dataset is better in accuracy and the generic framework can be used for any disease prediction. Also, it can be extended to develop more complex disease-specific chatbots.

C. Educational English Chatbots

In 2016, an interesting chatbot called SFITBOT was developed to answer frequently asked questions (FAQ) posted by students of St. Francis Institute of Technology (SFIT College) in Mumbai, India, without the need to browse a large number of webpages to obtain a response [15]. SFITBOT accepts a student’s question in natural language, generates a response by using pattern matching of the keywords in the query, and returns the response in the form of natural language to the user. SFITBOT was coded by using AIML and integrated with a website that stored student and college details in a MySQL database. SFITBOT provides the following additional features: it allows the student who did not receive an answer to his query to upload the query by using “upload query” function. Then, the teacher can upload his answer (including a PDF file or an image) to the system by using “upload file” function. Thus, SFITBOT has two types of users: students and teachers. Each of them has a special interface, and SFITBOT enables the teacher to interact with students. When the student logs into the system, a special interface will be displayed enabling the student to conduct the chat, with a button that enables the student to upload the query. When a student clicks on “upload query” button, a new interface will be displayed: the student can enter his query and send it to the teacher. If the teacher logs into the system, an interface will be displayed that enables the teacher to upload an answer (text, PDF file, or image). Unfortunately, it was not mentioned whether the system was tested by students.

D. Commercial English Chatbots

In 2015, Gupta et al. [16] designed an e-commerce website with a catalog of products: the user can browse it to choose the right product. Customers can search for the desired product by using traditional navigation. What distinguishes this site is that it can be integrated seamlessly with the chatbot. Hence, a chatbot was developed to help the customer decide on the right product. The chatbot can interact with the customer by using natural language: offering suggestions and asking questions. Technically, the e-commerce website was built by using HTML/CSS with PHP and a MySQL database to store details about products. The chatbot used Rivescript, which is a simple scripting language for giving intelligence to chatbots, to retrieve appropriate responses to user inputs.

E. Arabic Chatbots

In 2004, B.Abu Shawar and E. Atwell [17] adapted the Java program that was developed by them in 2003 [18] to produce Arabic AIML files that were extracted from Qur’an corpus. They visualized the data contents of Qur’an via chatting. The Java program, based on a machine learning approach, was developed to convert a corpus into AIML format to retrain ALICE [7]. It was able to create different AIML files by using different training corpora in different languages. Also, it was tested by using different corpora in many languages, such as English Dialogue Diversity Corpus and Afrikaans corpus [18]. The Java program generates two files: atomic file and default file. The atomic file includes the same questions and answers as they appeared in the corpus. The patterns in the atomic file represent the questions, while the templates represent the answers.

Since it cannot be guaranteed that the user will enter the same question as in ALICE knowledge base, the default file was built based on the first word and the most significant word approaches. The first word acts as a question classifier, while the most significant word is the least frequent word. For example, if the user asks, “What is your name,” the word “What” refers to the question classifier, and the most significant word is “name,” because it is the least frequent word. Hence, the answer will be based on the word “name” [19]. As known, the Qur’an is the holy book of Islam and it is a non-conversational text, so we can guarantee there are no overlaps, or breaks. They generated 76,404 AIML categories from Qur’an. Therefore, when a user types a word in chat (for example, "أَحَدٌ "), the system replies with a list of all ayyas that contain this word (for example, "قُلْ هُوَ اللَّهُ أَحَدٌ","وَلا يُوثِقُ وَثَاقَهُ أَحَدٌ") with soora’s title and ayya’s number. In addition, if the user writes the complete ayya, the system will reply with the next ayya in soora, soora's title, and ayya's number. The researchers tested the chatting dialogue with Qur’an. The system was able to accept the user’s input written in Arabic and reply with an appropriate ayya. In addition, they proved that their machine learning approach can deal with different corpora, regardless if it is conversational or not [17].

In 2011, Bayan Abu Shawar [19] suggested a new way to access Arabic web question answering (QA) corpus by using ALICE chatbot [7]. She modified the Java program that was used with Qur’an [17] and extended the FAQs chatbot that was generated for English and Spanish languages in [20] to include Arabic QA. She built a small corpus of 412 Arabic QA collected from different websites to cover topics such as mother’s and pregnancy issues, teeth care issues, fasting and health-related issues, blood tests such as cholesterol, diabetes, and blood charity issues. After she collected the Arabic QA from different websites, she modified the Java program as follows: for the atomic file, she removed punctuation and unnecessary symbols and inserted each question as a pattern and each answer as a template. For the default file, she extracted two most significant words that had a lower frequency. Moreover, she added different categories to increase the chance of finding answers. Next, she evaluated the system, and 93% of answers generated by the system were correct. Because the Arabic language has many characteristics, changing the question to another form is possible; therefore, this affects the answer and may lead to no answer.

In 2016, the first Arabic dialect chatbot was developed that uses the Egyptian Arabic dialect in the conversation; it is called BOTTA [21]. It represents a female character that converses with users for entertainment. BOTTA was developed by using AIML and was launched by using Pandorabots platform [22]. BOTTA’s knowledge base consists of AIML files, set files, and map files. In AIML files, BOTTA stores categories for several themes (such as greetings, gender, and nationality). The set files contain lists of word and phrases under one theme (for example, countries and months); thus, BOTTA can use these lists to generate general knowledge that helps BOTTA to continue the conversation and entertain users. The map files are used to relate words to words or phrases; hence, the map files contain lists of key–value pairs. For example, if a user enters his/her name, BOTTA searches for the name in names.set. If she finds it, she checks name2gender.map to determine the corresponding gender and generate responses according to the gender. Additionally, to overcome the inconsistent spelling variations of certain characters in Arabic, BOTTA used orthographic normalization. For example, in BOTTA every Alif-Maqsura “ى” was changed to “ ي“.

To evaluate BOTTA’s effectiveness, the researchers asked three native Arabic speakers (two of them were native Egyptian Arabic speakers and one was a Levantine Arabic speaker) to conduct a chat with BOTTA. All speakers agreed that BOTTA was amusing and entertaining, and they wanted to spend more time chatting. However, they noticed that sometimes she repeated the phrases and made out-of-context statements. Accordingly, they suggested to let BOTTA talk about herself more, let her ask the user more questions, and introduce new topics to lead the conversation. Despite the similarity in the idea between BOTTA and Nabiha, there are some differences between them. BOTTA represents a female character that speaks Egyptian Arabic dialect and is dedicated to entertainment. Meanwhile, Nabiha chatbot represents a female character that speaks the Saudi Arabic dialect and is dedicated to serving IT students. Another difference is that BOTTA is published on the Pandorabots platform only, while Nabiha is published on the Pandorabots platform and is integrated with an Android application, Twitter, and web-based platforms.

In 2018, AlHumoud et al. [23] presented a state-of-the-art review of research on Arabic chatbots. They examined twelve Arabic chatbots and classified them into two categories based on the conversation interaction type: text and speech. In each category, they classified the chatbots based on the implementation technique (either pattern matching or AIML approach), the length and domain of the conversation, and the dataset model. They concluded that research on Arabic chatbots is rare; the techniques used in implementing Arabic chatbots are pattern matching and AIML, and all available work is retrieval-based. This is owing to the complexity of the Arabic language and the lack of available resources to train the learning model.

F. Chatbots for Other Languages

In 2016 [24], a chatbot that uses Indonesian conversational patterns to conduct a human–computer conversation was developed. The knowledge of this chatbot was stored in the database, and the chat patterns were stored in MySQL relational database management system (RDBMS) tables as pattern–template. In addition, they used structured query language (SQL) to execute pattern-matching operation. For the pattern-matching process, they used bigram method to calculate sentence-similarity scores. The chatbot consists of two components: core (or RDBMS) and interface. The core contains a database (tables that store data) and an interpreter (a program that has sets of functions and procedures necessary to apply pattern matching). The interface is a standalone application that was implemented by using Pascal and Java programming languages. Moreover, they applied several tests to verify the application and found that the bigram method is applicable to other languages (not only Indonesian) with some constraints.

III. Methodology

We created our chatbot in five stages. First, we collected data from sources used by students, which contained their opinions, complaints, etc. Second, we built a dialogue corpus of files containing text in the Saudi dialect. Third, we generated AIML files by using a program that converts the readable text from the corpus into AIML format. Finally, we launched our chatbot on the Pandorabots platform and subsequently integrated it with Android, Twitter, and the Web. Bellow, we demonstrate the details of these stages.

A. Data Collection and Generation

To collect and prepare our data, we performed the following steps: data collection, data preprocessing, and data classification.

1) Data collection: We collected 248 inputs/outputs from the KSU IT students’ accounts in Askme.com. Three of them were personal accounts, while the other was a general account for the IT department. We have chosen to collect our data from Askme, because it contained students’ opinions, complaints, most asked questions, conversations of students among themselves, etc. Thus, we obtained a general idea of the conversations that a student can have with the chatbot. In addition, we extracted additional information from the KSU website and faculty members’ websites.

2) Data preprocessing: We preprocessed the collected data by removing the personal data, such as students’ names, personal opinions, or personal experiences. In addition, we removed the AIML-reserved symbols such as <, >, *, ^, #, _ .

3) Data classification: We classified the collected data into several text files. Some of them were related to the courses, while the others were related to the general rules and information related to the academic rules of KSU.

B. Bulding a Corpus

From the classified data, we built a dialogue corpus that contains several text files. For example, the greetings file contains greetings in the Saudi dialect plus some questions and sentences that enable the chatbot to learn basic information about the user such as name and undergraduate level (the study plan at the KSU IT department is composed of 8 levels). Our aim for building this corpus is to make it available for researchers working on Saudi chatbots, because some files are generic and can be used for any other chatbot (such as greetings file). In addition, we built a knowledge base of questions and answers about the IT department, which can be reused by other researchers.

C. Generating AIML Files

We developed two Java programs to convert the readable text into a chatbot format (AIML format). The first Java program converts the readable text file from the dialogue corpus to AIML format—specifically, to the basic formats (category, pattern, template); we added the remaining formats on the Pandorabots Playground. Hence, it learns from the dialogue corpus to generate AIML files and represent the patterns and templates underlying these dialogues, to make the chatbot behave like a human. The second Java program converts a single sentence (input/output) into AIML format (pattern/template).

D. Launching Chatbot on the Pandorabots Platform

After creating the AIML files that have the atomic category only through Java programs, we uploaded these files on the Pandorabots platform. In addition, we created other AIML files on the Pandorabots platform that have default and recursive categories.

E. Integrating AIML Chatbot with Android and Third-Party Platforms such as Twitter and Web

To facilitate access to Nabiha, we integrated our chatbot with an Android application, created a Twitter account for Nabiha, and developed a website for Nabiha. Section 4 presents the details of launching the chatbot on the Pandorabots platform. Fig. 1 presents the system's methodology.

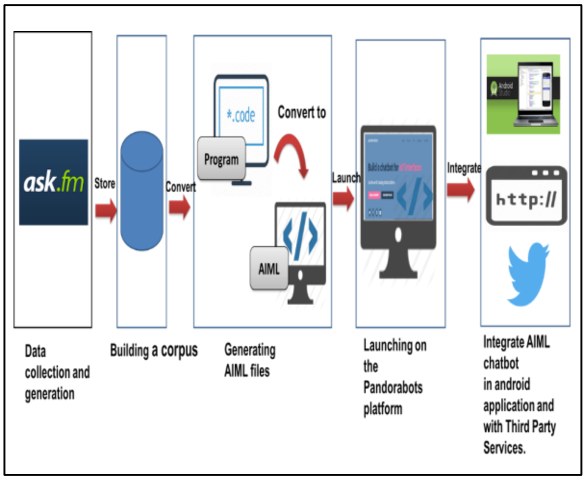

ask.fm
Store
Convert
.code
Convert to
Program
AIML
Launch
Integrate
http://
Data collection and generation
Building a corpus
Generating AIML files
Launching on the Pandorabots platform
Integrate AIML chatbot in android application and with Third Party Services.

Fig. 1. System Methodology.

IV. System Implementation and Integration

The implementation and integration process passed through several stages. First, we launched the chatbot on the Pandorabots platform. At this stage, we uploaded our AIML files that were created by Java programs and then we used the AIML editor that was provided by "Pandorabots Playground" to create other AIML files with different AIML tags and categories. Afterwards, we published Nabiha on the Pandorabots platform. Moreover, to make Nabiha available on different platforms, we integrated Nabiha with an Android application and deployed it on Twitter and on the web. Next, we describe the details of these stages.

A. Lanching Chatbot on the Pandorabots Platform

After we generated AIML files from our corpus, we launched it on the Pandorabots platform (Pb). As we mentioned before, Pandorabots platform is a web service for building and deploying chatbots. It provides a "Playground," which is an integrated development environment for building chatbots loaded with features and tools, and it also supports the concept of "Artificial Intelligence as a Service" (AIaaS) by providing a representational state transfer (RESTful) API for integrating artificially intelligent chatbots into applications and third-party tools. Thus, we created an account in Pandorabots Playground, and we used the Playground as an AIML editor. In addition, one of the features offered by Pandorabots Playground is the possibility to talk to the bot on its platform before deploying it, to enable the developer to make adjustments to his/her files. When we talked to our bot, we were surprised that it only responds to sentences that were stored in AIML files. For example, when we asked Nabiha about course Math106, whether it is hard or not (" بسألك عن ريض ١٠٦ هل هي صعبة و لا سهلة "), she did not have an answer, although there is a pattern similar to our question stored in Nabiha's knowledge base. This is because she was limited to responses to the patterns that were exactly as those stored in the knowledge base. For this reason, we regenerated the files by using keyword matching approach: for each file, we have one copy from our corpus and another based on keyword matching approach. In addition, to apply keyword matching approach in AIML language, we used the Wildcards, which is often used to provide an answer if there is no suitable category that can be matched to the user's input.

B. Talk to Nabiha Chatbot on the Pandorabots Platform

The Pandorabots platform allows the chatbot's developers to talk with their bot through "Train" tab and "Clubhouse" tab. The "Train" tab is an interactive development environment; we used it to talk with our bot and test each created category, to ensure that it is working correctly. Moreover, the Pandorabots platform allows the developers to publish their bot in "Clubhouse" that provides a chat user interface, thus, allowing other users to talk with their bots.

C. Pandorabots API

After we created and uploaded Nabiha's AIML files and talked to her on the Pandorabots platform, we registered in the "Dev portal" then received a private app_id and user_key to create a bot and upload bot files. The following four steps were required to integrate and deploy a chatbot:

1) Create a bot.
2) Upload bot files.
3) Compile bot.
4) Talk to your bot.

Additionally, we used Pandorabots API to manage our bot and integrate it with an Android application and deploy it with third-party Twitter and web-based platforms.

As we mentioned previously, Pandorabots platform provides AIaaS by offering a RESTful API, which is an API that uses HTTP requests to GET, PUT, POST, and DELETE data, and it can access Pandorabots hosting platform. Fig. 2 shows the strategy of communication between Pb RESTful API, Pb hostname, and developer. In addition, Pandorabots provides a command line interface (CLI) to allow the developers to manage his/her bot from the command line.

D. Integrate and Deploy Nabiha Chatbot

After we created bot in Dev portal and uploaded Nabiha files to the bot through RESTful API and Pb CLI, we integrated the bot with the Android application and deployed it on Twitter and on the web.

1) Nabiha chatbot in the Android application: To integrate our bot with the Android application, we followed these steps:

a) Create a chat interface in Android Studio by using Material ListView.

b) We used pb RESTful API and sent an HTTP request with POST method to the Pb hosting to retrieve data.

(Request.Method.POST,"https://aiaas.pandorabots.com/talk/app_id/botname")

c) The response from the bot is a JSON object. We store it in a JSON array and display it in the chat interface.

Fig. 3 shows the interface of the Nabiha Android application.

2) Nabiha chatbot on twitter: In addition, to make Nabiha act as a human, we made it available on Twitter by following these steps:

a) We created a Twitter user account for Nabiha.

b) We created a Twitter application on the Twitter website and obtained API keys (consumer key, consumer secret, access token, and access token secret).

c) We installed Pandorabots Python SDK (PbPython), which can interact with the Pandorabots API.

d) We installed Python library Tweepy to access Twitter API.

e) We downloaded file twitter_bot.py from GitHub and adapted it to be compatible with our bot.

f) We used the command “python twitter_bot.py --continuous true” in the command line to make the program run continuously.

3) Nabiha chatbot on the web: To make Nabiha available on the web, we created a website to allow students to talk with Nabiha and deployed it using the following steps:

a) We registered on Heroku cloud application platform, which is a “platform as a service” that enables us to deploy, run, and manage applications.

b) We deployed our bot on Heroku by using pb-html package.

c) We developed a website to access Nabiha and talk with her.

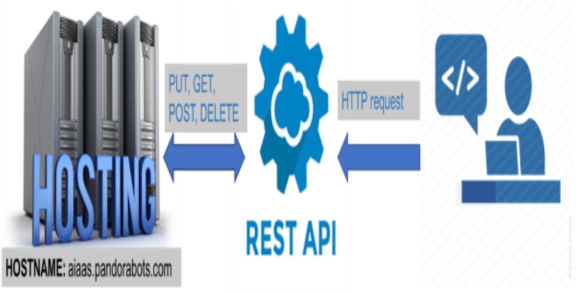


Fig. 2. Communication between Pb RESTful API, Pb Hostname, and Developer.

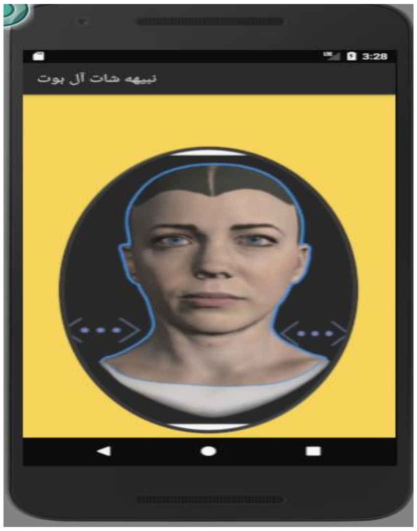


Fig. 3. Nabiha's Android Application.

V. Evaluation

In April 2017, Radziwill and Benton [25] proposed a methodology to assess the quality of chatbots; they described several quality issues and attributes that help in quality assessment of chatbots based on ISO 9241. We used these attributes to evaluate Nabiha chatbot [25]. ISO 9241 is a collection of international standards related to human–computer interaction and usability [26]. ISO 9241 defines usability as follows: “software is usable when it allows the user to execute his task effectively, efficiently, and with satisfaction in the specified context of use.” [26]. Thus, the three important attributes are effectiveness, efficiency, and satisfaction. Effectiveness refers to how well the system meets the user’s goals, i.e., it indicates accuracy and completeness. Efficiency refers to how well the resources are consumed to achieve the user’s goals. Satisfaction refers to user satisfaction: how they feel about using the system.

To ensure the usability of Nabiha chatbot, we asked 13 students to use the Nabiha chatbot and give their feedback by answering a questionnaire. We focused on testing Android and web platforms only, because Twitter’s text area only allows a limited number of characters.

In the questionnaire, we tried to ask questions that reflect effectiveness, efficiency, and satisfaction categories and quality attributes. To evaluate effectiveness, we focused on performance. There are several quality attributes related to performance, such as robustness to unexpected input and avoiding inappropriate utterances. To evaluate efficiency, we focused on functionality and humanity and their quality attributes, such as linguistic accuracy of the outputs and the ability to respond to specific questions. To evaluate satisfaction, we focused on affect, ethics, behavior, and accessibility and their quality attributes such as providing greetings, conveying personality and respect, inclusion, and preservation of dignity. Thus, we divided our questions into several categories:

- The methodology of Nabiha in conducting conversation.

- Quality of information provided by Nabiha.
- Nabiha’s abilities.
- Conversation with Nabiha in general.

VI. Results

After testing the usability of our chatbot in different categories, we obtained the results shown in Table I.

To ensure that Nabiha acts as a human and to simulate human behavior in conducting a conversation, we asked the students to imagine Nabiha’s identity. Many of them imagined that Nabiha would be an employee in the IT department, and some other participants imagined that Nabiha would be a graduate student who has a good knowledge and background. Only one person said Nabiha is a robot. Therefore, that means Nabiha somewhat succeeded in simulating a human conversation.

During the usability test with students, Nabiha was unable to reply to some sentences that existed in Nabiha’s knowledge base. For example, when a student asked, “how are you?” ("كيفك؟ "), Nabiha did not reply, although there is a category containing " كيفك " in Nabiha’s knowledge base.

```
<category>
<pattern># كيفك #<pattern>
<template>الحمدلله بخير أنتي كيفك؟</template>
</category>
```

The reason was the student typed a punctuation mark “?” without leaving a space between the word and mark. Thus, Nabiha interpreted "كيفك؟ " as a single word and searched for a category that contained this word. However, we solved this problem and considered similar cases.

To conclude this section, after performing usability testing of Nabiha chatbot and based on the questionnaire results, we can say the results of the first experiment with the Nabiha chatbot were somewhat acceptable. However, Nabiha still needs to be improved by increasing the dataset, although the dataset contained 1104 categories. The reason is the difficulty of the Arabic language, especially the Saudi dialect, and the lack of guarantee of a certain manner of asking questions. Additionally, we should find a way to solve the problem of HTML tags and should minimize some sentences to be compatible with the size of Twitter’s text area.

TABLE. I. Usability Testing Results

Category	Question	Results			
The methodology of Nabiha in conducting conversation.	How was the experience of your communication with Nabiha?	Easy 13/13 (100%)	Difficult —		
	Have you been welcomed?	Yes 13/ 13 (100%)	No —		
	Have Nabiha’s expressions been nice and pleasant?	Yes 12/13 (92.31%)	No 1/13 (7.69%)		
	Have Nabiha’s replies been polite and respectful?	Yes 13/13 (100%)	No —		
Quality of information provided by Nabiha.	How would you evaluate the level of information you got from Nabiha?	Bad —	Acceptable 3/13 (23.08%)	Good 4/13 (30.77%)	Excellent 6/13 (46.15%)
	Was she accurate in giving answers?	Not accurate 2/13 (15.38%)	Somewhat accurate 7/13 (53.85%)	Very precise 4/13 (30.77%)	
	Have Nabiha's answers been appropriate to the context of the user’s request / sentence?	Not appropriate 1/13 (7.69%)	Suitable 8/13 (61.54 %)	Very suitable 4/13 (30.77%)	
Nabiha's abilities.	How would you evaluate the language of Nabiha (linguistic accuracy, noting that Nabiha speaks in the dialect of Saudi Arabia)?	Bad —	Acceptable 12/13 (92.31%)		
	Was she able to continue the conversation on a particular topic (for example, a course)?	Yes 11/13 (84.62%)	No 2/13 (15.38%)		
	Did Nabiha interact with your conversation?	Yes 13/13 (100%)	No —		
	In general, was Nabiha able to conduct a conversation?	Bad —	Good 9/13 (69.23%)	Excellent 4/13 (30.77%)	
	Has Nabiha replied promptly?	Slow 1/13 (7.69%)	Fast 12/13 (92.31%)		
Conversation with Nabiha in general.	How was chatting with Nabiha?	Bad —	Acceptable 3/13 (23.08%)	Good 6/13 (46.15%)	Excellent 4/13 (30.77%)
	The conversation in general has been…	Bad —	Somewhat satisfactory 7/13 (53.85%)	Convincing 6/13 (46.15%)	
	How satisfied are you with Nabiha?	Dissatisfied —	Somewhat satisfied 9/13 (69.23%)	Totally satisfied 4/13 (30.77%)	
	Will you chat Nabiha again?	Yes 12/13 (92.31%)	No 1/13 (7.69%)		

VII. Conclusion

Communication with computers using natural language is one of the most interesting and exciting interactions for the user. In this paper, we present a new Arabic dialect chatbot called Nabiha, which is dedicated to serve the students of the IT department at King Saud University, enabling them to conduct a conversation using the Saudi dialect. Our chatbot is the first chatbot that uses the Saudi dialect and combines entertainment and usefulness. It enables the student to spend a fun time when she feels stressed and gives good advice when the students require counseling. The results of the first user experience evaluation of Nabiha chatbot was somewhat acceptable. However, there are some limitations and problems: the dataset should be increased; we need to solve the problem of HTML tags and deal with the limitations of Twitter's text area. Hence, we intend to make these improvements in the future. In addition, we plan to make Nabiha available on other platforms such as WhatsApp, Skype, iOS, and instant messages.

Acknowledgment

This research project was supported by a grant from the "Research Center of the Female Scientific and Medical Colleges", Deanship of Scientific Research, King Saud University.

References

[1] N. J. Nilsson, Artificial intelligence: A modern approach: Stuart Russell and Peter Norvig,(Prentice Hall, Englewood Cliffs, NJ, 1995);, vol. 11. Elsevier, 1996.

[2] A. Shaikh, G. Phalke, P. Patil, S. Bhosale, and J. Raghatwan, "A Survey On Chatbot Conversational Systems," Int. J. Eng. Sci., vol. 3117, 2016.

[3] B. Setiaji and F. W. Wibowo, "Chatbot Using A Knowledge in Database," presented at the Intelligent Systems, Modelling and Simulation (ISMS), 2016 7th International Conference, Bangkok, Thailand, 2016, pp. 72–77.

[4] B. A. Shawar and E. Atwell, "Chatbots: are they really useful?," in LDV Forum, 2007, vol. 22, pp. 29–49.

[5] J. Weizenbaum, "ELIZA— a computer program for the study of natural language communication between man and machine," Commun. ACM, vol. 26, no. 1, pp. 23–28, 1983.

[6] L. H. Young and W. Xiang, "A literature review of the implementation of Dialogue based Natural Language Chatbot and their practical applications."

[7] A. van Woudenberg, "A Chatbot Dialogue Manager-Chatbots and Dialogue Systems: A Hybrid Approach," Open Universiteit Nederland, 2014.

[8] D. Bacon, "SETL fot Internet Data Processing," 2005.

[9] S. M. Freudenberger, J. T. Schwartz, and M. Sharir, "Experience with the SETL optimizer," ACM Trans. Program. Lang. Syst. TOPLAS, vol. 5, no. 1, pp. 26–45, 1983.

[10] N. Polatidis, "Chatbot for admissions," ArXiv Prepr. ArXiv14086762, 2014.

[11] A. S. Lokman and J. M. Zain, "An architectural design of Virtual Dietitian (ViDi) for diabetic patients," in Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on, 2009, pp. 408–411.

[12] B. E. V. Comendador, B. M. B. Francisco, J. S. Medenilla, and S. Mae, "Pharmabot: a pediatric generic medicine consultant chatbot," J. Autom. Control Eng. Vol, vol. 3, no. 2, 2015.

[13] N. Belgaumwala, "Chatbot: A Virtual Medical Assistant," Int. J. Res. Appl. Sci. Eng. Technol., vol. 7, no. VI, p. 9, Jun. 2019.

[14] M. Bali, S. Mohanty, S. Chatterjee, M. Sarma, and R. Puravankara, "Diabot: A Predictive Medical Chatbot using Ensemble Learning."

[15] Rodrigues, Natalina, Salgaonkar,Samiksha, and Rego, Natasha, "SFITBOT," Int. J. Comput. Sci. Inf. Technol., vol. 7, no. 2, pp. 954–956, 2016.

[16] S. Gupta, D. Borkar, C. De Mello, and S. Patil, "An E-Commerce Website based Chatbot," vol. 6, no. 2, pp. 1483–1485, 2015.

[17] A. Shawar and E. S. Atwell, "An Arabic chatbot giving answers from the Qur'an," in Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles, 2004, vol. 2, pp. 197–202.

[18] B. A. Shawar and E. Atwell, "Machine Learning from dialogue corpora to generate chatbots," Expert Update J., vol. 6, no. 3, pp. 25–29, 2003.

[19] B. A. Shawar, "A Chatbot as a Natural Web Interface to Arabic Web QA.," iJET, vol. 6, no. 1, pp. 37–43, 2011.

[20] A. Shawar, E. Atwell, and A. Roberts, "FAQchat as in Information Retrieval system," in Human Language Technologies as a Challenge for Computer Science and Linguistics: Proceedings of the 2nd Language and Technology Conference, 2005, pp. 274–278.

[21] D. A. Ali and N. Habash, "Botta: An Arabic Dialect Chatbot," pp. 208–212, 2016.

[22] Kathrin Haag, "The Learning, Teaching and Web Services Division (LTW) Teacher Bot Project Pandorabots Playground Documentation," Univ. Edinb. Home, pp. 1–20, May 2016.

[23] S. AlHumoud, A. Al Wazrah, and W. Aldamegh, "Arabic Chatbots: A Survey," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 8, pp. 535–541, 2018.

[24] B. Setiaji and F. W. Wibowo, "Chatbot Using a Knowledge in Database: Human-to-Machine Conversation Modeling," in Intelligent Systems, Modelling and Simulation (ISMS), 2016 7th International Conference on, 2016, pp. 72–77.

[25] N. M. Radziwill and M. C. Benton, "Evaluating Quality of Chatbots and Intelligent Conversational Agents," ArXiv Prepr. ArXiv170404579, 2017.

[26] A. Abran, A. Khelifi, W. Suryn, and A. Seffah, "Usability meanings and interpretations in ISO standards," Softw. Qual. J., vol. 11, no. 4, pp. 325–338, 2003.

Personality Classification from Online Text using Machine Learning Approach

Alam Sher Khan[1], Hussain Ahmad[2], Muhammad Zubair Asghar[3]*
Furqan Khan Saddozai[4], Areeba Arif[5], Hassan Ali Khalid[6]
Institute of Computing and Information Technology
Gomal University, D.I. Khan, Pakistan

Abstract—**Personality refer to the distinctive set of characteristics of a person that effect their habits, behaviour's, attitude and pattern of thoughts. Text available on Social Networking sites provide an opportunity to recognize individual's personality traits automatically. In this proposed work, Machine Learning Technique, XGBoost classifier is used to predict four personality traits based on Myers- Briggs Type Indicator (MBTI) model, namely Introversion-Extroversion(I-E), iNtuition-Sensing(N-S), Feeling-Thinking(F-T) and Judging-Perceiving(J-P) from input text. Publically available benchmark dataset from Kaggle is used in experiments. The skewness of the dataset is the main issue associated with the prior work, which is minimized by applying Re-sampling technique namely random over-sampling, resulting in better performance. For more exploration of the personality from text, pre-processing techniques including tokenization, word stemming, stop words elimination and feature selection using TF IDF are also exploited. This work provides the basis for developing a personality identification system which could assist organization for recruiting and selecting appropriate personnel and to improve their business by knowing the personality and preferences of their customers. The results obtained by all classifiers across all personality traits is good enough, however, the performance of XGBoost classifier is outstanding by achieving more than 99% precision and accuracy for different traits.**

Keywords—Personality recognition; re-sampling; machine learning; XGBoost; class imbalanced; MBTI; social networks

I. INTRODUCTION

Personality of a person encircles every aspect of life. It describes the pattern of thinking, feeling and characteristics that predict and describe an individual's behaviour and also influences daily life activities including emotions, preference, motives and health [1].

The increasing use of Social Networking Sites, such as Twitter and Facebook have propelled the online community to share ideas, sentiments, opinions, and emotions with each other; reflecting their attitude, behaviour and personality. Obviously, a solid connection exists between individual's temperament and the behaviour they show on social networks in the form of comments or tweets [2].

Nowadays personality recognition from social networking sites has attracted the attention of researchers for developing automatic personality recognition systems. The core philosophy of such applications is based on the different personality models, like Big Five Factor Personality Model [3], Myers- Briggs Type Indicator (MBTI) [4], and DiSC Assessment [5].

The existing works on personality recognition from social media text is based on supervised machine learning techniques applied on benchmarks dataset [6], [7], [8]. However, the major issue associated with the aforementioned studies is the skewness of the datasets, i.e. presence of imbalanced classes with respect to different personality traits. This issue mainly contributes to the performance degradation of personality recognition system.

To address the aforementioned issue different techniques are available for minimizing the skewness of the dataset, like Over-sampling, Under-sampling and hybrid-sampling [9]. Such techniques, when applied on the imbalanced datasets in different domain, have shown promising performance in terms of improved accuracy, recall, precision, and F1-score [10].

In this work, a machine learning technique, namely, XGBoost is applied on the benchmark personality recognition dataset to classify the text into different personality traits such as Introversion-Extroversion(I-E), iNtuition-Sensing(N-S), Feeling-Thinking(F-T) and Judging-Perceiving(J-P). Furthermore, to improve the performance of the system, resampling technique [11] is also utilized for minimizing the skewness of the dataset.

A. Problem Statement

Predicting personality from online text is a growing trend for researchers. Sufficient work has already been carried out on predicting personality from the input text [6, 7, 8].

However, more work is required to be carried out for the performance improvement of the existing personality recognition system, which in most of the cases arises due to presence of imbalanced classes of personality traits. In the proposed work. A dataset balancing technique, called re-sampling is used for balancing the personality recognition dataset, which may result in improved performance.

B. Research Questions

RQ.1: How to apply supervised machine learning technique, namely XGBoost classifier for classifying personality traits from the input text?

RQ.2: How to apply a class balancing technique on the imbalanced classes of personality traits for performance improvement and what is the efficiency of the proposed technique w.r.t other machine learning techniques?

*Corresponding Author

RQ.3: What is the efficiency of the proposed technique with respect to other baseline methods?

C. Aims and Objective

1) Aim: The aim of this work is to classify the personality traits of a user from the input text by applying supervised machine learning technique namely XGBoost classifier on the benchmark dataset of MBTI personality. This work is the enhancement of the prior work performed by [6].

2) Objectives

a) Applying machine learning technique namely XGBoost classifier for personality traits recognition from the input text.

b) Applying re-sampling technique on the imbalanced classes of personality traits for improving the performance of proposed system.

c) Evaluating the performance of proposed model with respect to other machine learning techniques and base line methods.

D. Significance of Study

Personality is distinctive way of thinking, behaving and feeling. Personality plays a key role in someone's orientation in various things like books, social media sites, music and movies [12].

The proposed work on personality recognition is an enhancement of the work performed by [6]. Proposed work is significant due to the following reasons: (i) performance of the existing study is not efficient due to skewness, which will be addressed in this proposed work by applying re-sampling technique on the imbalanced dataset, (ii) proposed work also provide a basis for developing state of the art applications for personality recognition, which could assist organization for recruiting and selecting appropriate personnel and to improve their business by taking into account the personality and preferences of their customers.

II. Related Work

A review of literature pertaining to personality recognition from text is presented here in this section. The literature studies of this work is categorized into four sub groups, namely, i) Supervised learning techniques, ii) Un-supervised machine learning techniques, iii) Semi-supervised machine learning techniques and, iv) Deep learning techniques.

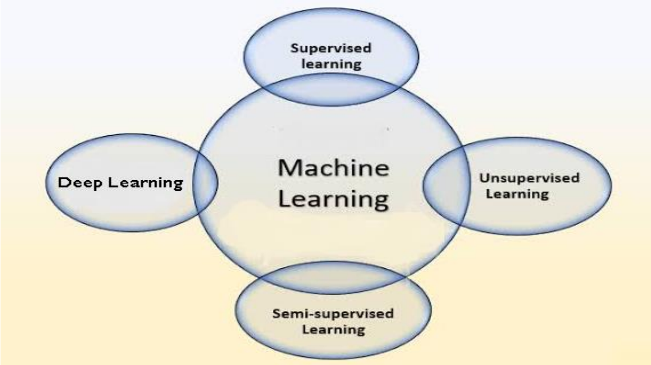


Fig. 1. Categorization Sketch of Literature Review.

Fig. 1 depicts the classification sketch of the literature review on personality recognition from text.

A. Supervised Learning Technique

These supervised learning algorithms are comprised of unlabeled data/ variables which is to be determined from labelled data, also called independent variables. The studies given below are based on supervised learning methodologies.

A system is proposed by [6] for analysing social media posts/ tweets of a person and produce personality profile accordingly. The work mainly emphasizes on data collection, pre-processing methods and machine learning algorithm for prediction. The feature vectors are constructed using different feature selection techniques such as Emolex, LIWC and TF/IDF, etc. The obtained feature vectors are used during training and testing of different kinds of machine learning algorithms, like Neural Net, Naïve Bayes and SVM. However, SVM with all feature vectors achieved best accuracy across all dimensions of Myers-Briggs Type Indicator (MBTI) types. Further enhancement can be made by incorporating more state of the art techniques.

MBTI dataset, introduced in [7] for personality prediction, which is derived from Reddit social media network. A rich set of features are extracted, and benchmark models are evaluated for personality prediction. The classification is performed using SVM, Logistic Regression, and (MLP). The classifier using all linguistic features together outperformed across all MBTI dimensions. However, further experimentation is required on more models for achieving more robust results. The major limitation is that the number of words in the posts are very large, which sometimes don't predict the personality accurately.

To predict personality from tweets, [8] proposed a model using 1.2 Million tweets, which are annotated with MBTI type for personality and gender prediction. Logistic regression model is used to predict four dimensions of MBTI. Binary word n-gram is used as a feature selection. This work showed improvement in I-E and T-F dimensions but no improvements in S-N and even slightly drop for P-J. In terms of personality prediction, linguistic features produce far better results. Incorporating enhanced dataset may improve performance.

A system was developed to recognize user personality using Big Five Factor personality model from tweets posted in English and Indonesian language [13]. Different classifiers are applied on the MyPersonality dataset. The accuracy achieved by Naive Bayes(NB) is 60%, which is better than the accuracy of KNN (58%) and SVM (59%).Although this work did not improve the accuracy of previous research (61%) yet achieved the goal of predicting the personality from Twitter-based messages. Using extended dataset and implementing semantic approach, may improve the results.

Personality assessment/ classification system based on Big5 Model was proposed for Bahasa Indonesian tweets [14]. Assessment is made on user's words choice. The machine learning classifiers, namely, SVM and XGBoost, are implemented on different parameters like existence of (n_gram minimum and n_gram weighted), removal of stop words and using LDA. XGBoost performed better than the SVM under the same data and same parameter setting. Limited dataset of only 359 instances for training and testing is the main drawback of their work.

Automatic identification of Big Five Factor Personality Model was proposed by [15] using individual status text from Facebook. Various techniques like Multinomial NB, Logestic Regression (LR) and SMO for SVM are used for personality classification. However, MNB outperformed other methods. Incorporating feature selection and more classifiers, may enhance the performance.

Personality profiling based on different social networks such as Twitter, Instagram and Foursquare performed by [16]. Multisource large dataset, namely NUS-MSS, is utilized for three different geographical regions. The data is evaluated for an average accuracy using different machine learning classifiers. When the different data sources are concatenated in one feature vector, the classification performance is improved by more than 17%. Available dataset may be enriched from multi (SNS) by user's cross posting for better performance.

The performance of different ML classifiers are analysed to assess the student's personality based on their Twitter profiles by considering only Extraversion trait of Big 5 [17]. Different machine learning algorithms like Naïve Bayes, Simple logistic, SMO, JRip, OneR, ZeroR, J48, Random Forest, Random Tree, and AdaBoostM1, are applied in WEKA platform. The efficiency of the classifiers is evaluated in terms of correctly classified instances, time taken, and F-Measures, etc. OneR algorithm of rules classifier show best performance among all, producing 84% classification accuracy. In future, all dimensions of Big5 can be considered for evaluation to get more insight.

The performance of different classifier is evaluated by [18] using MBTI model to predict user's personality from the online text. Various ML classifiers, namely Naïve Bayes, SVM, LR and Random Forest, are used for estimation. Logistic Regression received a 66.5% accuracy for all MBTI types, which is further improved by parameter tuning. Results may further be improved by using XGBoost algorithm, which remained winner of most Kaggle and other data science competitions.

The oversampling and undersampling techniques are compared by [11] for imbalance dataset. Classification perform poorly when applied on imbalanced classes of dataset. There are three approaches (data level, algorithmic level and hybrid) that are widely used for solving class imbalance problem. Data level method is experimented in this study and result of Over-sampling method (SMOTE) is better than under-sampling technique (RUS). More re-sampling techniques need to be evaluated in future.

Authors in [19] briefly discussed and explained the early research for the classification of personality from text, carried out on various social networking sites, such as Twitter, Blogger, Facebook and YouTube on the available datasets. The methods, features, tools and results are also evaluated. Unavailability of datasets, lack of identification of features in certain languages, and difficulty in identifying the requisite pre-processing methods, are the issues to be tackled. These issues can be addressed by developing methods for non-English language, introducing more accurate machine learning algorithms, implementing other personality models, and including more feature selection for pre-processing of data.

Twitter user's profiles are used for accurate classification of their personality traits using Big5 model [20]. Total 50 subjects with 2000 tweets per user are assessed for prediction. Users content are analysed using two psycholinguistic tools, namely LIWC and MRC. The performance evaluation is carried out using two regression models, namely ZeroR and GP. Results for "openness" and "agreeableness" traits are similar as that of previous work, but less efficient results are shown for other traits. Extended dataset may improve the results.

A connection has been established between the users of Twitter and their personality traits based on Big5 model [21]. Due to inaccessibility of original tweets, user's personality is predicted on three parameters that are publicly available in their profiles, namely (i) followers, (ii), following, and (iii) listed count. Regression analysis is performed using M5 rules with 10-fold cross validation. RMSE of predicted values against observed values is also measured. Results show that based on three counts, user's personality can be predicted accurately.

TwiSTy, a novel corpus of tweets for gender and personality prediction has been presented by [22] using MBTI type Indicator. It covers six languages, namely Dutch, German, French, Italian, Portuguese and Spanish. Linear SVM is used as classifier and results are also tested on Logistic Regression. Binary features for character and word (n-gram) are utilized. It outperformed for gender prediction. For personality prediction, it outperformed other techniques for two dimensions: I-E and T-F, but for S-N and J-P, this model did not show improvement. In future, the model can be trained enough to predict all four dimensions of MBTI efficiently.

The Table I represents the summaries of above cited studies for classification and prediction of user's personality using Supervised Machine Learning strategies.

B. *Unsupervised Learning Approach*

Unsupervised learning classifiers are using only unlabeled training data (Dependent Variables) without any equivalent output variables to be predicted or estimated.

The Twitter data was annotated by [23] for 12 different linguistic features and established a correlation between user's personality and writing style with different cross-region users and different devices. Users with more than one tweets are considered for evaluation. It was observed that Twitter users are secure, unbiased and introvert as compared to the users posting from iPhone, blackberry, ubersocial and Facebook platforms. More Twitter data for classification may enhance the efficiency of personality identification model.

TABLE I. PERSONALITY RECOGNITION BASED WORK USING SUPERVISED MACHINE LEARNING APPROACH

SNo	Research	Goals and objectives	Strategy/ Approach	Performance	Limitation and Future Work
1	Bharadwaj et al. (2018) [6]	Personality prediction from online text	SVM, Neural Net and Naïve Bayes TF-IDF, Emolex, LIWC and ConceptNet	SVM with all feature vectors achieved best accuracy across all dimensions of MBTI	Less weightage is given to the word's gravity. Incorporating more state-of-the-art techniques in future will yield better result.
2	Gjurković and Šnajder (2018) [7]	Personality classification of Reddit user's posts.	SVM, Logistic Regression and MLP with linguistic features	MLP using all linguistic features together outperform across all MBTI dimensions	Demographic data like age and gender is not considered Accuracy of T/F dichotomy may be improved in future.
3	Plank and Hovy (2015) [8]	Personality and gender prediction from tweets.	Logistic regression Model and Binary word n-gram is used as a feature selection.	Accuracy for personality prediction: I/E = 72.5% S/N = 77.5% T/F = 61.2 % J/P = 55.4%	A lot of Gap between general population personality types and this corpus personality types. Incorporating of enhanced dataset will improve the performance.
4	Pratama and Sarno (2015) [13]	To recognize user personality using Big-5 personality model from tweets posted in English and Indonesian language	Supervised • KNN • NB • SVM	Accuracy KNN = 58% NB = 60% SVM = 59%	Using extended dataset and implementing semantic approach, may improve the results.
5	Ong et al. (2017b) [14]	A personality assessment based on Big5 Model for Bahasa Indonesian tweets using user's words choice.	Supervised • XGBoost • SVM	Accuracy XGBoost = 97.99% SVM = 76.23%	Limited dataset of only 359 instances for training and testing is the main drawback of this work.
6	Alam et al. (2013) [15]	Automatic identification of Big Five Factor Personality Model using individual status text from Facebook	Multinomial NB, Logestic Regression (LR) and SMO for SVM are used for personality classification	MNB = 61.79% BLR = 58.34% SMO = 59.98% ›MNB outperformed other methods	Incorporating feature selection and more classifiers, may enhance the performance.
7	Buraya et al. (2017) [16]	Multisource large dataset, namely NUS-MSS, is utilized for personality profiling.	Supervised	By concatenating different data sources in one feature vector, the classification performance is improved by more than 17%.	In future the available dataset may be enriched from multi (SNS) by user's cross posting for better performance.
8	Ngatirin et al. (2016) [17]	Using different ML classifiers to assess the student's personality based on their Twitter profiles.	Naïve Bayes, Simple logistic, SMO, JRip, OneR, ZeroR, J48, Random Forest, Random Tree, and AdaBoostM1,	OneR with F1_Score = 0.837 outperform among all.	In future, all dimensions of Big5 can be considered for evaluation to get more insight.
9	Chaudhary et al. (2018) [18]	To predict user's personality from the online text using MBTI model.	Supervised learning methodology namely Naïve Bayes, SVM, LR and Random Forest, are used for estimation.	Accuracy NB = 55.89% LR = 66.59% SVM = 65.44%	Lower accuracy is due using traditional classifiers. Deep learning approach will definitely improve the performance.
10	Kaur and Gosain (2018) [11]	Comparing of oversampling and undersampling techniques for imbalance dataset.	Decision tree algorithm C4.5 is used.	Result of Over-sampling method (SMOTE) is better than under-sampling technique (RUS).	More re-sampling techniques need to be evaluated in future.
11	Ong et al. (2017a) [19]	Classification of personality from text, carried out using various social networking sites.	Survey paper using supervised learning approcah	Best result among all was attained by twitter with 91.9% accuracy using words frequency.	Unavailability of datasets, and lack of identification of features in certain languages, are the issues to be tackled. In future methods for non-English language may need to be developed.
12	Golbeck et al. (2011) [20]	User's Twitter profiles for accurate classification of their personality traits using Big5 model.	Two regression models, namely ZeroR and GP are used.	Accuracy Higher for Open = 75.5% Lower for Neuro =42.8%	Extended dataset may improve the results.

13	Quercia et al. (2011) [21]	To establish a connection between the users of Twitter and their personality traits based on Big5 model.	Regression using M5 rules with 10-fold cross validation.	RMSE: O = 0.69 C = 0.76 E = 0.88 A = 0.79 N = 0.85	In future user personality classification may be utilized in marketing and recommender system.
14	Verhoeven et al. (2016) [22]	To predict gender and personality from a novel corpus of tweets, namely TwiSTy.	SVM and logistic Regression along words n_grams features.	Ƒ_score I/E =77.78 S/N =79.21 T/F = 52.13 J/P = 47.01 For italic lang:	In future, the model can be trained enough to predict all four dimensions of MBTI efficiently.

The purpose of the study carried out by [24], is to scrutinize the group-based personality identification by utilizing unsupervised trait learning methodology. Adawalk technique is utilized in this survey. The outcomes portray that while considering Micro- Ƒ1 score, the achievement of adawalk is exceptional with somewhat 7% for wiki, 3% for Ƈora, and 8% for BlogCaṯlog. While utilizing SoCE personality corpus, 97.74% Macro-Ƒ1 score was achieved by this approach. The drawback of this work is that it entirely depends on TƑ -IDƑ strategy, additionally the created content systems are not an impersonation of genuine social and interpersonal network like retweeting systems. Large and increased dataset will definitely enhance the performance of the proposed work in future.

An unsupervised personality classification strategy was accomplished by [25] to highlight the matter that to how extent different personalities collaborate and behave on social media site Twitter. Linguistic and statistical characteristics are utilized by this work and then tested on data corpus elucidated with personality model using human judgment. System investigation anticipate that psychoneurotic users comments more than secure ones and tend to develop longer chain of interaction.

An Unsupervised Machine learning methodology, namely, Ḳ-Meańs was accomplished by [26] to recognize the network visitors' trait and personality. This proposed work is based on the quantifiable contents of the website. The obtained results portray that this strategy can be utilized to predict website and network visitors' personality traits, more accurately. Proposed system may be enhanced in future by adding more elements associated with websites and a greater number of websites for the better performance.

Author in [27] proposed a personality identification system using unsupervised approach based on Big-5 personality model. Different social media network sites are used for extraction and classification of user's traits. Linguistic features are exploited to build personality model. The system predict personality for an input text and achieved reasonable results. However, extended annotated corpus can boost the system's performance.

TABLE II. PERSONALITY RECOGNITION BASED WORK USING UN-SUPERVISED MACHINE LEARNING APPROACH

SNo	**Research**	**Goals and objectives**	**Strategy/ Approach**	**Outcome**	**Limitation and Future Work**
1	Celli (2011) [23]	Personality classification from individual's writing pattern	Un_supervised Score-based	Mean Accuracy =0.6651 and Mean validity= 0.6994	Additional Tweets for personality recognition may improve the accuracy of this proposed model.
2	Sun et al. (2019) [24]	group-based personality identification	Un_supervised Adȧwalk	97.74% (Macɾo-Ƒ1)	Large and increased dataset will definitely enhance the performance of the proposed work in future
3	Celli and Rossi, (2012) [25]	Impact of linguistic characteristics on personality traits.	Un-supervised Statistics-based	78.29% (Accurȧcy)	More tweets are needed for efficient investigation
4	Chishti and Sarrafzadeh (2015) [26]	To recognize the network visitors' trait and personality	Uń-supervised Ḳ-Mean	Ḳ=10 is accurate score	System may be enhanced in future by adding more elements associated with websites and a greater number of websites for better performance
5	Celli (2012) [27]	Impact of linguistic characteristics on personality traits using Big Five Model	Un_supervised Score-based	81.43% (Accuracy)	Extended annotated corpus can boost the system's performance
6	Arnoux et al. (2017) [28]	Developing personality model to predict individual's Big Five personality traits on much fewer data using twitter.	Word-Embedding	68.5% (Accuracy)	Findings of this method are based on English Twitter data, which may be extended to other languages

A model was proposed by [28] that requires eight times fewer data to predict individual's Big Five personality traits. GloVe Model is used as Word embedding to extract the words from user tweets. Firstly, the model is trained and then tested on given tweets. Further, the data is tested on three other combinations: (i) GloVe with RR, (ii) LIWC with GP, and (iii) 3-Gram with GP, and the proposed model performed better with an average correlation of 0.33 over the Big-5 traits, which is far better than the baseline method. Findings of this method are based on English Twitter data, which may be extended to other languages. Similarly, the performance of the model can be examined with small number of tweets.

The Table II illustrates the concise detail of above cited studies regarding user's personality and traits identification from textual data using un-supervised machine learning approach.

C. Semi-Supervised Learning Approach

The studies carried out by using the combination of linguistic and lexicon features, supervised machine learning methodologies and different feature selection algorithms are known as semi-supervised ML approaches. The following studies have utilized the semi-supervised and hybrid strategy.

Multilingual predictive model was proposed by [29], which identified user's personality traits, age and gender, based on their tweets. SGD classifier with n-gram features, is used for age and gender classification, while LIWC with regressor model (ERCC) is used for personality prediction. An average accuracy of 68.5% has been achieved for recognition of user's attributes in four different languages. However, author profiling can be enhanced by performing experiments in multiple languages.

A technique was devised to detect MBTI type personality traits from social media (Twitter) in Bahasa Indonesian language [30]. Among 142 respondents, 97 users are selected with an average 2500 tweets per user. WEKA is used for building classification and training set. Three approaches are used for prediction from training set. i) Machine Learning, ii) Lexicon-based, and iii) linguistic Rules driven. Among all, Naïve Bayes outperformed the comparing methods in terms of better accuracy and time. Its accuracy for I/E trait is 80% while for S/N, T/F and J/P, its accuracy is 60%. Lower accuracy on the part of linguistic rule-driven and lexicon-based, are due to limited corpus in Bhasha Indonesia. It is observed that by increasing the training data set, accuracy may get improved.

A technique proposed for personality prediction from social media-based text using word count [31]. It works for both MBTI and Big5 personality models using 8 different languages. Four kinds of labelled corpus both for Big5 and BMTI are used for conducting the experiments. In each corpus, 1000 most frequently used words are selected. Prediction accuracy for "openness" trait of Big5 is higher across all corpus, while for MBTI, prediction accuracy for S/N dimension is greater than other dichotomies. Using only word count for prediction is the main drawback of the proposed system, which may be covered by introducing different features selection and ML algorithms.

Detail of the above quoted studies regarding personality classification using Semi-supervised Machine Learning Approach are presented in Table III.

TABLE III. PERSONALITY RECOGNITION BASED WORK USING SEMI-SUPERVISED MACHINE LEARNING APPROACH

SNo	**Research**	**Goals and objectives**	**Strategy/ Approach**	**Performance**	**Limitation and Future Work**
1	Arroju et al. (2015) [29]	Multilingual predictive model is used to identify user's personality traits, age and gender, based on their tweets.	›SGD classifier with n-gram features. › LIWC with regressor model (ERCC)	Accuracy = 68.5%	Accuracy may be improved by using different personality model. Similarly, author profiling can be further enhanced by performing experiments in multiple languages.
2	Lukito et al. (2016) [30]	To recognize MBTI type personality traits from social media (Twitter) in Bahasa Indonesian language.	›Machine Learning, ›Lexicon-based, and ›linguistic Rules driven	I/E trait = 80% S/N, T/F and J/P accuracy is 60%	Lower accuracy is due to limited corpus in Bhasha Indonesia. By increasing the training data set, accuracy may get improved.
3	Alsadhan and Skillicorn (2017) [31]	Personality prediction from social media-based text using word count	Based on word count	Accuracy for "openness" trait of Big5 is higher, while for MBTI, accuracy for S/N dimension is greater than all other dichotomies.	Using only word count for prediction is the main drawback of the proposed system, which may be covered by introducing different features selection and ML algorithms.

D. Deep Learning Strategy

Deep learning is a subcategory of machine learning (ML) in artificial intelligence (AI), where machines may acquire knowledge and get experience by training without user's interaction to make decisions. Based on experiences and learning from unlabeled and unstructured corpus, deep learning performs tasks repeatedly and get improvement and tweaking in results after each iteration. The studies given below are in summarized form, showing the prior work performed in Deep learning.

A deep learning classifier was developed, which takes text/tweet as input and predict MBTI type of the author using MBTI dataset [32]. After applying different pre-processing techniques embedding layer is used, where all lemmatized words are mapped to form a dictionary. Different RNN layers are investigated, but LSTM performed better than GRU and simple RNN. While classifying user, its accuracy is 0.028 (.676 × .62 × .778 × .637), which is not good. The predictive efficiency of this work may be improved by increasing the number of posts per user. As the model is tested on real life example of Donald trump's 30,000 tweets, which correctly predict his actual MBTI type personality.

A model proposed by [33] that takes snippet of post or text as input and classify it into different personality traits, such as (INFP, ENTP, and ISJF, etc.). Different classification methods like Softmax as baseline, SVM, Naïve Bayes, and deep learning, are implemented for performance evaluation. SVM outperformed NB and softmax with 34% train 33% test accuracy, while Deep learning model shows more improvement with 40% train and 38% test accuracy. However, the accuracy is still low as it doesn't even achieve 50 percent.

Personality classification system is proposed by [34], to recognize the traits from online text using deep learning methodology. AttRCNN model was suggested for this study utilizing hierarchical approach, which is capable of learning complex and hidden semantic characteristics of user's textual contents. Results produced are very effective, proving that using deep and complex semantic features are far better than the baseline features.

A deep learning model was suggested by [1] to classify personality traits using Big Five personality model based on essay dataset. Convolutional Neural Network (CNN) is used for this work to detect personality traits from input essay. Different pre- processing techniques like word n-grams, sentence, word and document level filtration and extracting different features are performed for personality traits classification. "OPN" traits achieved higher accuracy of 62.68% by using different configuration of features and among all five traits. In future, more features need to be incorporated and LSTM recurrent network may be applied for better results.

Table IV represents the outline of the works regarding automatic personality recognition system using Deep learning methodology.

TABLE IV. PERSONALITY RECOGNITION BASED WORK USING DEEP LEARNING APPROACH

SNo	Research	Goals and objectives	Strategy/ Approach	Outcome	Limitation and Future Work
1	Hernandez and Scott (2017) [32]	To predict and classify people into their MBTI types using their online textual contents.	Deep Learning ›RNN ›LSTM ›GRU ›BiLSTM	Accuracy I/E= 67.6% S/N=62.0% T/F=77.8% J/P=63.7%	The predictive efficiency of this work may be improved by increasing the number of posts per user.
2	Cui, and Qi (2018) [33]	A model that takes snippet of post or text as input and classify it into different personality traits.	Deep Learning Multi-layer LSTM	Over all accuracy= 38% I/E= 89.51% S/N=89.84 % T/F=69.09 % J/P=69.37 %	In future more deep learning techniques with more word embedding features may be exploited. Using of unsupervised technique will also give better results.
3	Xue et al. (2018) [34]	To recognize the personality traits from online text using deep learning methodology.	Deep Learning using AttRCNN Approach	MAE OPN= 0.3577 CON= 0.4251 EXT= 0.4776 AGR= 0.3864 NEU= 0.4273	In future these deep and complex semantic features will be used as input of regression classifiers for more improvement in the performance.
4	Majumder et al. (2017) [1]	To classify personality traits using Big Five personality model based on essay dataset.	Deep Learning ›CNN	Accuracy OPN= 62.68% CON= 56.73% EXT= 58.09% AGR= 56.71% NEU= 59.38%	In future more features need to be incorporated and LSTM recurrent network may be applied for better results.

III. Methodology

The working procedure of this proposed system are as follows: (i) Data acquisition and re-sampling, (ii) Pre-Processing and feature selection, (iii) Text-based Personality classification using MBTI model, (iv)Applying XGBoost for personality classification, (v) Comparing the efficiency of XGBoost with other classifiers, (vi) Applying different evaluation metrics.

A. Dataset Collection and Re-sampling

The publically available benchmark dataset is acquired from Kaggle [6]. This data set is comprised of 8675 rows, where every row represents a unique user. Each user's last 50 social media posts are included along with that user's MBTI personality type (e.g. ENTP, ISJF). As a result, a labelled data set comprising of a total 422845 records, is obtained in the form of excerpt of text along with user's MBTI type. Table V describes the detail of acquired dataset.

1) Re-Sampling: As pointed out by [6], the original dataset is totally skewed and unevenly distributed among all four dichotomies, described as follows: **I/E Trait:** I=6664 and, E= 1996, **S/N Trait:** S= 7466 and N= 1194, **T/F Trait:** T= 4685 and F= 3975, **J/P Trait:** J= 5231 and P= 3429. Whenever, an algorithm is applied on skewed and unbalanced classified dataset, the outcome always diverge toward the sizeable class and the smaller classes are bypassed for prediction. This drawback of classification is known as class imbalance problem (CIP) [11].

Therefore, this sparsity is balancedby re-sampling technique [11]. As mentioned earlier, two traits are highly imbalanced, Data Level Re-sampling approach for class balancing is used [9]. This bridged the gap between each dichotomy traits and resulted in the efficient and predictable performance of the proposed system.

TABLE V. Detail of Dataset

Dataset Name	Description	Instances	Format	Default Task	Updated	Origin	Size	Creator
MBTI_kaggle	This dataset was acquired from Kaggle by using PersonalityCafe platform. The members of PerC have known MBTI personality type along their tweets or text. The dataset comprised of 8676 PerC members personality types.	8675	Text	Personality Prediction	2018	Kaggle	25 MB	Mitchell J

2) Data Level Re-Sampling Approach: Data manipulation sampling approaches focus on rescaling the training datasets for balancing all class instances. Two popular techniques of class resizing are over-sampling and under-sampling.

At the data level, the most famous methodologies are Oversampling and under sampling procedures. Oversampling is the way toward expanding the number of classes into the minority class. The least difficult oversampling is random oversampling, which basically duplicate minority instances to enhance the imbalance proportion.This duplication of minority class enhancement really improved the performance of machine learning classifier for efficient personality traits prediction [11].

Under samplingapproach is used to level class distribution by indiscriminately removing or deleting majority class instances. This process is continued till the majority and minority class occurrences are balanced out.

As illustrated in Fig. 2, the data level sampling-based methodologies including over-sampling and under-sampling have gotten exceptionalconsiderations to counter the impact of imbalanced datasets [35].

3) Training and Testing Data: In this proposed system, the data is divided into Training, Testing and Validation dataset. Mostly two datasets are required, one for building the model while the other dataset is needed to measure the performance of the model. Here training and validation are used for building the model, while Testing step is used to measure the performance of the proposed model [36]. Table VI shows the sample tweets from training dataset, while Table VII represents the sample of test data tweets.

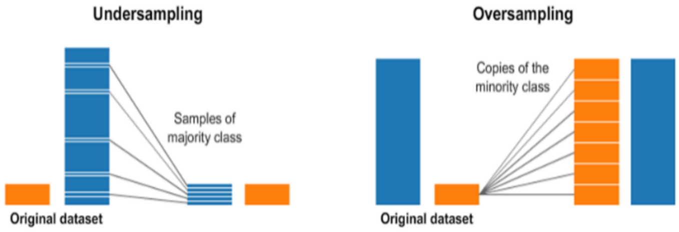


Fig. 2. Class Balancing using Undersampling and Oversampling.

TABLE VI. Sample Tweets from Training Dataset

Personality Type	Tweets
ISTP	'I'm only a mystery to myself.
INTP	Of course, to which I say I know; that's my blessing and my curse.
INFJ	Hello ENFJ7. Sorry to hear of your distress.
ENTP	'I'm finding the lack of me in these posts very alarming.
ENTJ	Lol. Its not like our views were unsolicited. What a victim.
INFP	That more or less finds myself in agreement, honey cookie.
ESTP	Most things hands on. For me, music. I'm very tactile. I like to write too.

TABLE VII. SAMPLE TWEETS FROM TEST DATASET

Personality Type Type	Tweets
ENFP	Patience is a virtue. So proud that you guys are still together.
ISFJ	We are always willing to help those in need
ENTJ	I'm scared of failure, but also throwing up...take that for what you will.
INFP	That would be the best description for what I usually am.
ENFJ	You're right. Not sure why I didn't think of that before hahah
ESTP	I have 0 friends. I don't trust anybody.

At the point when the dataset is divided into training data, validating data and testing data, it utilizes just a portion of dataset and it is clear that training on minor data instances the model won't behave better and overrate the testing error rate of algorithm to set on the whole dataset.

To address this problem a cross-validation technique will be used.

4) Cross-validation: It is a statistical methodology that perform splitting of data into subgroups, training on the subset of data and utilize the other subset of data to assess the model's authentication.

Cross validation comprises of the following steps:

- Split the dataset into two subsets.
- Reserve one subset data.
- Train the model on the other subset of data.
- Using the reserve subset of data for validation (test) purpose, if the model exhibits better on validation set, it shows the effectiveness of the proposed model.

Cross validation is utilized for the algorithm's predictive performance estimation.

a) K fold cross validation: This strategy includes haphazardly partitioning the data into k subsets of almost even size. The initial fold is reserved for testing and all the remaining k-1 subsets of data are used for training the model. This process is continued until each Cross-validation fold (of k iteration) have been used as the testing set.

This procedure is repeated kth times; therefore, the Mean Square Error also obtained k times (from Mean Square Error-1 to kth Mean Square Error). So, k-fold Cross Validation error is calculated by taking mean of the Mean Square Error over Kfolds. Fig. 3, explain the working procedure of K-Fold cross validation.

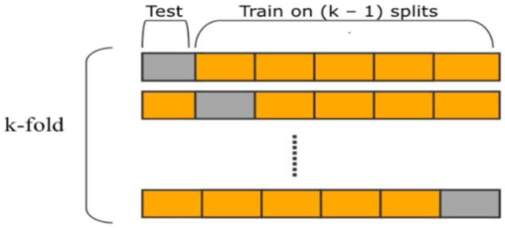


Fig. 3. K-Fold Cross Validation Working Procedure.

Algorithm 1. Dividing the Data set in Train and Test sets.

```
#Division of Data in training and testing sets:
Assign [] to Ӿ↔Ŧrȧin
Assign [] to Ȳ↔Ŧrȧin
Assign [] to Ӿ↔Ŧest
Assign [] to Ȳ↔Ŧest
Allocate Ŧest⟶ Ṩize to 20% of n̲
Assign RNDM (0, n̲ -1, Ŧest→Ṩize) to ŦINDICES
For Ị = 0 t̲o n̲-1
          Assign [] to Ŧemp
          Ƒor each ⱲỌRD in Ŧf-Idf [i]
                    Append (If-Idf [i][WORD]) to temp
          END FOR
If Ị in ŦINDICES then
          Ăppend (ŤEMṖ) to Ӿ↔Ŧest
          Ăppend (tweet [i][ Ị]) to Ȳ↔Ŧest
Ēlse
          Ăppend ŤEMṖ) to Ӿ↔Ŧrȧin
          Ăppend ŤEMṖ) to Ȳ↔Ŧrȧin
ȨƝĐ ỊƑ
ȨƝĐ ƑOR
```

B. Preprocessing and Feature Selection

Different pre-processing techniques and various feature selection are exploited, for more exploration of the personality from text. These techniques include tokenization, removal of URLs, User mentions and Hash tag, word stemming, stop words elimination and feature selection using TF IDF [28] and [32].

1) Preprocessing: The following preprocessing steps on mbti_kaggle dataset are applied before classification, acquired from the [37] work.

a) Tokenization: Tokenization is the procedure where words are divided into the small fractions of text. For this reason, Python-based NLTK tokenizer is utilized.

b) Dropping Stop Word: Stop words don't reveal any idea or information. A python code is executed to take out these words utilizing a pre-defined words inventory. For instance, "the", "is", "an" and so on are called stop words.

c) Word stemming: It is a text normalization technique. Word stemming is used to reduce the inflection in words to their root form. Stem words are produced by eliminating the pre-fix or suffix used with root words.

2) Feature Selection: The following feature selection steps are accomplished using different machine learning classifiers.

a) CountVectorizer: Using machine learning algorithms, it cannot execute text or document directly, rather it may firs be converted into matrix of numbers. This conversion of text document into numbers vector is called tokens.

The count vector is a well-known encoding technique to make word vector for a given document. CountVectorizer takes what's known as the Bag of Words approach. Each message or document is divided into tokens and the number of times every token happens in a message is counted.

CountVectorizer perform the following tasks:

- It tokenizes the whole text document.
- It constitutes a dictionary of defined words.

- It encodes the new document using known word vocabulary.

b) Term Frequency: It represents the weight of a word that how much a word or term occurs in a document.

c) Inverse document Frequency: It is also a weighting scheme that describe the common word representation in the whole document.

d) Term Frequency Inverse Document Frequency: The TF-IDF score is useful in adjusting the weight between most regular or general words and less ordinarily utilized words. Term frequency figures the frequency of every token in the tweet however, this frequency is balanced by frequency of that token in the entire dataset. TF-IDF value shows the significance of a token in a tweet of whole dataset [38].

This measure is significant in light of the fact that it describes the significance of a term, rather than the customary frequency sum [39].

Feature engineering module pseudocode is illustrated in the following Algorithm 2.

Algori̲thm2. Stepwise procedure for Ƒeature Engineering

```
# CountVectorizer
Assign [] to CVectorizer
Ƒor        Ēach tweet in Post Ɗo
           ƑorĒach word in tweet Ɗo
                      Assign Ɗict [word] to Ɗict [Ⱳord] +1
           ĒndƑor
           CVectorizer. Ăppend (Ɗict)
           Ässign 0 to Ɗict
ĒndƑor
ŦermƑrequency
Assign CVectorizer to TƑ
Assign 0 to ɌOẄ
Ⱳhile (ɌOẄ <= Ɲ-1) Ɗo
           Assign SUM (CVectorizer [row].values) to Nwords
           For Each Word in CVectorizer [row]
           Assign CVectorizer[W]/Nwords to TF [W]
           ĒndƑor
ⱲhileĒnd
# ŦF/ ƊƑ
# ƖDƑ Ꞓalculation
Assign [] to ƖDƑ
While (Till the existence of ɌOẄ in TƑ) Do
Assign [] to t̲emp
Ⱳhile (Till the existence of word in ɌOẄ) DO
           Assign 0 to Count
           Ƒor i fɍom 0 to Ɲ-1 Ɗo
                      IƑ TƑ [Count][Ẅord]>0 Ŧhen
                         Count← Count+1
                      Ēnd IƑ
           ĒndƑor
                      Ässign LÖG (Ɲ/Count) to Ťemp [Ẅord]
           ⱲhileĒnd
           IDF. Äppend (ŦEMṖ)
ⱲhileĒnd
# ŦF/ IDƑ
Assign 0 to TƑ -ƖDƑ
ƑOR I Ƒrom 0 to Ɲ-1 ƊÖ
           Assign []  to  ŦĒMṖ
           ƑorĒach Ẅ in TƑ [i], ƖDƑ [i]
                      ŦĒMṖ [W]= TƑ [i][Ẅ]*ƖDƑ[i][ Ẅ]
           ĒndƑor
           Äppend (ŦEMṖ) to TƑ -ƖDƑ
ĒndƑor
```

C. Text-based Personality Classification Using MBTI Model

In this proposed work, supervised learning approach is used for personality prediction. The model will take snippet of post or text as an input and will predict and produce personality trait (I-E, N-S, T-F, J-P) according to the scanned text. Mayers-Briggs Type Indicator is used for classification and prediction [4]. This model categorize an individual into 16 different personality types based on four dimensions, namely, (i) *Attitude →E̲xtroversion vs I̲ntroversion:* this dimension defines that how an individual focuses their energy and attention, whether get motivated externally from other people's judgement and perception, or motivated by their inner thoughts, (ii) *Information →S̲ensing vs iN̲tuition (S/N):* this aspect illustrates that how people perceive information and observant(S), relying on their five senses and solid observation, while intuitive type individuals prefer creativity over constancy and believe in their guts, (iii) *Decision →T̲hinking vs F̲eeling (T/F):* a person with Thinking aspect, always exhibit logical behaviour in their decisions, while feeling individuals are empathic and give priority to emotions over logic, (iv) *Tactics →J̲udging vs P̲erceiving (J/P):* this dichotomy describes an individual approach towards work, decision-making and planning. Judging ones are highly organized in their thoughts. They prefer planning over spontaneity. Perceiving individuals have spontaneous and instinctive nature. They keep all their options open and good at improvising opportunities [40].

D. Working Procedure of the System for Personality Traits Prediction

As depicted in Fig. 4, first, the proposed model is trained by giving both labelled data (MBTI type) and text (in the form of tweets). After training the model, it is evaluated for efficiency. For better prediction, the dataset will be split into three phases (training phase, validating phase and testing phase). The validating step will reduce overfitting of data.

The mbti_kaggle dataset is available in two columns, namely, (i) type and (ii) posts. By type it means 16 MBTI personality types, such as INTP, ENTJ and INFJ, etc. As we are interested in MBTI traits rather than types, therefore we through python coding added four new columns to the original dataset for the purpose of traits determination. As a result, the new modified dataset will look like as given bellow in Table VIII.

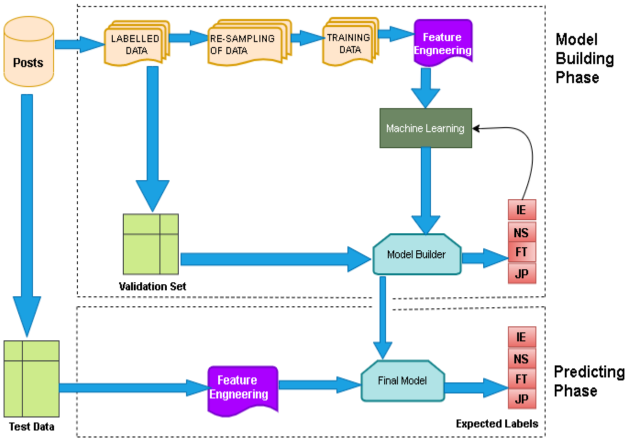


Fig. 4. Working Procedure of the System.

TABLE VIII. SAMPLE OF DATASET USED FOR EXPERIMENT

Type	Posts	I/E	S/N	F/T	J/P
ENTP	I'm scared of failure, but also throwing up...take that for what you will.	1	0	0	1
INFJ	Just a funny comment from my side. A bit serious maybe. If you don't care about the functions	0	0	1	0
INFP	I need a date with an INTJ! God dammit. Opps, wrong thread. lol	0	0	1	1

Algoriṯhm 3: Pseudo code of the entire System

```
Inṗut:                  Ṣet of tweets from mbti_kaggle dataset saved in CSV
format
Output:                 Classification of input text into personality traits
Personality Traits:                 [“I_E”, “S-N”, “F-T”, “J-P”]
ML-Classifier:                      [ “X̄Gboost”]
Stop-word List:                     [There, it, on, into, under…….]
Start
//Inputting Snippet of Text
Assign Dataset text of post to Text
#Pre-processing steps.
#Tokenization/segmentation
            Assign Tokenize(text) to Token
# Dropping of stop words
            Set Post_text to Drop_stopwords(tokens)
#punctuation
# data set splitting into train/teṣt
            Set Ӿ↔Ŧrȧin, Ȳ↔Ŧrȧin, Ӿ↔Ŧest, Ȳ↔Ŧest to Ṣplit (post_text,
ṯest-ṣize=20%)
# counterVectorizer(Post_Text)
#Application of tf‣ idf
#Classifier implementation
Set Model to MLClassifier
ÄssignMödel: fit(Ӿ↔Ŧrȧin, Ȳ↔Ŧrȧin) to Classification
Set Classification to Mödel: fit(Ӿ↔Ŧrȧin, Ȳ↔Ŧrȧin)

#Traits Ṗrediction
Assign Classification: Ṗrediction (Ӿ↔Ŧest) to Ṗrediction
Set Trait_Prediction to Classification: Ṗrediction (Ӿ↔Ŧest)
#Accuracy
Set Accuracy to Accuracy (Trait_Ṗrediction, Ȳ↔Ŧest)
#Recall score
Set Recall to Recall (Trait_Prediction, Ȳ↔Ŧest)
#Precision score
Set Precision to Precsion (Trait_Prediction, Ȳ↔Ŧest)
#F1‣ score
Set F1‣ score to F1‣ score (Trait_Prediction, Ȳ↔Ŧest)
Assign (Accuracy, Re_call, Precession, F1‣ score) to Personality Traits
Return (Personality Traits)
```

E. Applying XGBoost for Personality Classification

XGBoost belongs to the family of Gradient Boosting. It is used to handle classification and regression issues that make a prediction/ forecast from a set of weak decision trees.

Although work has been performed on personality assessment using supervised machine learning approaches [13, 17]. Here state of the art Algorithm XGBoost with optimized parameters is used for MBTI personality assessment [41]. XGBoost classifier is good on producing better accuracy as compared to other machine learning algorithms [41, 42]. The proposed work is the first attempt to predict personality from text using XGBoost as classifier and MBTI as personality model.

Algorithm 4: XGBoost Working Procedure

```
Data: Dataset and Hyperparameters
Initialize
for k = 1,2, ………, M do
Calculate $gk =$ ;
Calculate $hk =$ ;
Determine the structure by choosing splits with maximized gain
$\mathcal{A} =$
Determine the leaf weights = ;
Determine the base learner $b(x) =$ ;
Add trees $f_k(x) = f_{k-1}(x) + b(x)$;
end
Result: $f(x) =$
```

F. Comparing the Efficiency of XGBoost with other Classifiers

The overall prediction performance and efficiency of the proposed system has examined by applying other supervised machine learning classifiers. This comparison illustrates a true picture of the performance of this proposed classifier, namely XGBoost, as compared to the other machine learning algorithms and baseline methods regarding personality prediction capability from the input text [13].

G. Evaluation Metrics

The evaluation metrics, such as accuracy, precision, recall and f-measure, describe the performance of a model. Therefore, different evaluation metrics has been used to check the overall efficiency of predictive model.

Algorithm 5: Pseudo code of the Performance Evaluation

```
# Ṗerformance
ŦC
Äccuracy ↔ŦC/Ɲ2
TṖ  CÖUNŦ (Ṗredicṯion = Ṗosiṯive ÄNḊ Ỵ↔Ŧest =Ṗosiṯive)
TƝ ↔ CÖUNŦ (Ṗredicṯion =Ɲegaṯive ÄNḊ Ỵ↔Ŧest = Ɲegaṯive)
ƑṖ ↔ CÖUNŦ ((Ṗredicṯion = Ṗosiṯive ÄNḊ Ỵ↔Ŧest = Ɲegaṯive)
ƑN ↔ CÖUNŦ (Ṗredicṯion = Ɲegaṯive ÄNḊ Ỵ↔Ŧest = Ṗosiṯive)
Ṗrecisioń ↔TṖ / (TṖ + ƑṖ)
Recȧll ↔TṖ / (TṖ + ƑN)
CFM ↔ []
CƑM [‘TṖ’] ↔TṖ
CƑM [‘FN’] ↔ƑN
CFM[‘FP’] ↔ƑP
CƑM [‘ŦN’] ↔ŦN
```

IV. RESULTS AND DISCUSSIONS

This chapter presents a set of results which are produced from the proposed system by systematically answering the raised research questions.

A. Answer to RQ.1

To answer to RQ1: “How to apply supervised machine learning technique, namely XGBoost classifier for classifying personality traits from the input text?”, the supervised machine learning technique, XGBoost classifier is applied to predict MBTI personality traits from excerpt of text. Fine-tuned parameter setting for XGBoost is presented in Table IX.

Table X shows the results of XGBoost classifier with default parameter settings.

It is clear from Table XI that increasing or decreasing the values of different parameters for XGBoost classifier, has huge effect on the text classification results.

B. *Answer to RQ.2*

While addressing RQ2: "How to apply a class balancing technique on the imbalanced classes of personality traits for performance improvement and What is the efficiency of the proposed technique w.r.t other machine learning techniques?", An imbalanced dataset is considered first. Imbalanced dataset can be defined as a distribution problem arises in classification where the number of instances in each class is not equally divided.

Whenever, an algorithm is applied on skewed and unbalanced classified dataset, the outcome always diverge toward the sizeable class and the smaller classes are bypassed for prediction. This drawback of classification is known as class imbalance problem [11].

Therefore, it is attempted to balance this sparsity by re-sampling technique [11]. As two traits are highly imbalanced, therefore Data Level Re-sampling approach is used for class balancing [9].

TABLE IX. Parameter Setting for XGBoost

Parameters	**Description**
Learning_rate = 0.03	It describes the effect of weighting of adding more trees to the boosting model.
Colsample_bytree = 0.4	It corresponds to the fraction of features (columns) that will be used to train each tree.
Scale-pos_weight = 1	It controls the balance between negative and positive classes.
Subsample = 0.8	Subsample ratio of the training instance. Setting it to 0.5 means that XGBoost randomly collects half of the data instances to grow trees. This prevents overfitting.
Objective = 'binary:logistic',	It returns predicted probability for binary classification.
n_estimators = 1000	It represents the number of decision trees in XGBoost classifier.
Reg_alpha = 0.3	L1 regularization encourages sparsity (meaning pulling weights to 0).
Max-depth = 10	It represents the size (depth) of each decision tree in the model. Over fitting can be controlled using this parameter.
Gamma = 10	Its purpose is to control complexity. It represents that how much loss has to be reduced. It prevents overfittings.

TABLE X. Results of XGBoost without Parameter Settings

	Metrics	**I-E**	**S-N**	**F-T**	**J-P**
No Parameter setting	**Accuracy**	87.04	92.32	89.00	85.85
	Recall	81.44	81.75	87.70	89.16
	Accuracy	91.59	68.98	91.65	87.80
	F1_Score	86.22	74.82	89.92	88.47

TABLE XI. Results of XGBoost with different Parametersettings

	Metrics	**I-E**	**S-N**	**F-T**	**J-P**
learning_rate: 0.01 n_estimators: 1000 max_depth: 5 subsample: 0.8 colsample_bytree: 1 gamma: 1 Objective = 'binary:logistic' Reg_alpha = 0.3 Scale-pos_weight = 1	Accuracy	93.10	96.70	92.32	90.88
	Recall	89.56	96.24	92.07	94.24
	Precession	96.32	97.14	93.64	90.91
	F1_Score	92.82	96.68	92.85	92.55
learning_rate: 0.01 n_estimators: 1000 max_depth: 6 subsample: 0.8 colsample_bytree: 1 gamma: 1 Objective = 'binary:logistic' Reg_alpha = 0.3 Scale-pos_weight = 1	Accuracy	95.51	97.61	93.15	91.79
	Recall	93.39	97.21	92.91	94.77
	Precession	97.47	98.00	94.37	91.81
	F1_Score	95.39	97.60	93.64	93.27
learning_rate: 0.01 n_estimators: 500 max_depth: 6 subsample: 0.8 colsample_bytree: 1 gamma: 1 Objective = 'binary:logistic' Reg_alpha = 0.3 Scale-pos_weight = 1	Accuracy	90.95	94.51	91.20	89.84
	Recall	85.78	91.98	90.28	95.23
	Precession	95.48	96.88	93.28	88.69
	F1_Score	90.37	94.37	91.75	91.84
learning_rate: 0.01 n_estimators: 1000 max_depth: 10 subsample: 0.8 colsample_bytree: 1 gamma: 1 Objective = 'binary:logistic' Reg_alpha = 0.3 Scale-pos_weight = 1	Accuracy	99.37	99.92	94.55	95.53
	Recall	97.16	100	89.96	92.66
	Precession	100	99.50	100	100
	F1_Score	98.56	99.75	94.72	96.19

In this section the overall comparison of predicting personality traits is presented using all evaluation metrics to determine the performance of different classifiers. Results are reported in Table XII.

Different classifiers are applied over same mbti_kaggle dataset using Re-sampling technique and without Re-sampling technique. Results reported in Table XII depict that XGBoost obtained the highest score using all four-evaluation metrics and across all the MBTI personality dimensions, when imbalance dataset is experimented. However, Naïve Bayes and Random Forest on imbalance dataset, performed poorly. So, it is concluded from this experiment that applying classifiers on skewed data is not producing good results.

On the other hand, when different classifiers are tested over resampled dataset, an improved result is obtained for all dimensions over all classifiers.

The most accurate and precise algorithm for this proposed work is XGBoost. It got excellent results for all traits using all metrics. XGBoost obtained maximum accuracy (99.92%) for S/N trait. Its results are highest for all four dimensions and across all metrics.

1) Why our Class balancing technique is better: By applying class balancing technique results for all evaluation metrics and for all four personality traits are high and better than base line work. In this dataset two dimensions I/E and S/N are highly imbalanced, therefore a class balance technique is used for better prediction performance.

TABLE XII. COMPARISON OF DIFFERENT CLASSIFIERS PERFORMANCE USING RE-SAMPLE DATASET AND IMBALANCE DATASET

Classifier	Metrics	Without Re-sampling				With Re-Sampling			
		I-E	**S-N**	**F-T**	**J-P**	**I-E**	**S-N**	**F-T**	**J-P**
KNN	Accuracy	77.02	86.65	60.11	59.31	86.90	81.44	73.45	81.52
	Recall	20.34	15.5	89.89	77.17	86.44	98.00	93.82	89.19
	Precession	45.74	32.29	58.64	63.70	65.74	42.79	68.69	81.74
	F1_Score	28.16	20.94	70.98	69.79	74.51	59.57	79.31	85.30
Decision Tree	Accuracy	78.69	82.01	70.42	69.33	99.34	99.93	90.85	91.30
	Recall	53.31	38.25	71.94	75.11	97.00	99.50	83.14	85.72
	Precession	51.84	36.34	73.11	74.68	100	100	100	100
	F1_Score	52.56	37.27	72.52	74.89	98.48	99.75	90.79	92.31
Random Forest	Accuracy	77.93	86.03	74.,89	64.90	98.36	99.45	82.15	91.62
	Recall	00	0	84.49	97.7	92.59	98.94	74.07	86.24
	Precession	1	0	73.31	63.84	100	100	100	100
	F1_Score	00	0	78.50	77.22	96.15	99.44	85.10	92.61
MLP	Accuracy	83.83	88.40	83.41	75.86	99.27	99.93	94.52	92.18
	Recall	40.37	22.0	84.68	86.46	96.69	99.59	89.90	87.91
	Precession	83.83	88.40	83.41	75.86	100	100	100	81.906
	F1_Score	40.37	22.0	84.68	86.46	98.32	99.75	88.89	93.14
SVM	Accuracy	85.54	88.68	85.02	78.62	95.94	98.08	92.63	91.37
	Recall	43.69	22.75	85.64	90.36	91.32	97.00	89.45	91.11
	Precession	82.93	85.84	86.59	78.01	90.28	90.02	96.73	94.53
	F1_Score	57.23	35.96	86.12	83.74	90.69	93.38	92.95	92.79
MNB	Accuracy	77.86	86.03	54.63	60.92	79.32	88.82	84.04	60.11
	Recall	0	0	99.93	100	6.78	20.25	73.18	100
	Precession	0	0	54.47	60.91	97.73	98.78	96.68	60.11
	F1_Score	0	0	70.51	75.71	12.66	33.61	83.25	75.09
XGboost	Accuracy	86.52	89.21	83.16	80.82	99.37	99.92	94.55	95.53
	Recall	52.68	31.5	84.04	89.90	97.16	100	89.96	92.66
	Precession	79.52	78.26	84.80	80.78	100	99.50	100	100
	F1_Score	63.38	44.92	84.42	85.10	98.56	99.75	94.72	96.19
Logistic Reg	Accuracy	82.47	86.48	84.32	76.63	92.80	96.09	88.96	88.44
	Recall	25.86	4.5	86.35	93.52	85.33	90.25	85.28	92.14
	Precession	83.67	78.26	84.99	74.57	82.72	83.18	93.90	89.23
	F1_Score	39.51	8.5	85.66	82.98	84.01	86.57	89.34	90.66
SGD	Accuracy	85.26	90.29	85.19	79.36	94.31	97.42	91.86	90.99
	Recall	41.64	40.5	85.71	90.82	91.64	95.50	87.52	89.39
	Precession	83.54	80.19	86.83	78.61	84.08	87.21	97.21	95.53
	F1_Score	55.58	53.82	86.27	84.28	87.70	91.17	92.11	92.36

KNN classifier gives overall low performance, however its Recall for I/E and F/T is a little bit high.

The outcome of Decision Tree algorithm for I/E and S/N traits is better than F/T and J/P traits.

Random Forest gives highest for all traits. However, for J/P lowest Recall is obtained.

Logistic Regression classifier produced tremendous result for all traits, but again for J/P traits accuracy and Precision are not up to the mark.

The results obtained by applying Naïve Bays classifier is comparatively better for I/E and S/N traits.

Support Vector Machine when tested on the given dataset it gives better and balance results in respect to all traits. SGD Classifier showing remarkable performance for all four personality traits.

MLP classifier achieved outstanding results for all four traits using four metrics.

XGBoost classifier has proven to be very good for classification problems. The results obtained using XGBoost is very balance in respect to all personality traits

C. Answer to RQ.3

To answer RQ3**:** "What is the efficiency of the proposed technique with respect to other baseline methods." This proposed model is compared with two baseline methods [6, 7].

Classification performed by [6] for personality prediction using same mbti_kaggle dataset by applying three classifiers namely, (i) SVM, (ii) MLP and (iii) Naïve Bayes and got accuracy upto 88.4%. Due to imbalance data the result of [6] is not up to the mark. The results show that SVM in collaboration with LIWC and TF-IDF feature vectors gave accurate prediction score for all four traits, while MLP with all features Vectors got maximum accuracy score for S/N trait (90.45%) however its result for J/P trait is lower. Naïve bays also perform well for I/E and S/N traits but its performance for T/F and J/P is very poor. The reason behind better accuracy for I/E and S/N dimensions and least performance for T/F and J/P is due to class imbalance problem.

A very large dataset MBTI9k acquired from reddit is used for personality prediction [7]. The emphasis of this work is to extract features and linguistic properties of different words and then these features are used to train various machine leaning models such as Logistic Regression, SVM and MLP. Classifiers using integration of all features together (LR_all and MLP_all) obtained better results for all traits. The overall worst results using all classifiers obtained for the T/F dichotomy. The major limitation of this work is that the number of words in each post are very large, which lead to a little bit lower performance on the part of all classifiers.

1) Proposed Work: In this proposed system, the same dataset is used as experimented by [6], However re-sampling technique is applied over it, and hence obtained results in respect of all personality traits are very good, especially XGBoost achieved the best score across all dimensions and all traits as compared to previous work. It is observed that the mbti_kaggle dataset is very skewed, therefore when oversampling technique is applied the output is far better than all previous works. Up to 99% accuracy for I/E and S/N traits are achieved using XGBoost classifier, while Bharadwaj [6], got 88% maximum accuracy for S/N trait. Similarly, for T/F and J/P proposed work results are promising and obtained 94.55% accuracy for T/F and 95.53% accuracy for J/P dimension using XGBoost. While in previous work MLP classifier achieved accuracy of 54.1% for T/F and 61.8% for J/P dimension. Therefore, it is clear that by using resampling technique excellent and improved results are obtained for all four dimensions. The results reported in Table XIII, describe the comparison of proposed work with the baseline method.

2) XGBoost with Outstanding Performance: XGBoost belongs to the family of Gradient Boosting is a machine learning technique used for classification and regression problems that produces a prediction from an ensemble of weak decision trees.

The main reason of using this algorithm is its accuracy, speed, efficiency, and feasibility. It's a linear model and a tree learning algorithm that does parallel computations on a single machine. It also has extra features for doing cross validation and computing feature importance.

TABLE XIII. COMPARISON OF XGBOOST WITH BASELINE TECHNIQUE

Study	Technique	Dataset	Classifier	Obtained Results				
				Metrics	I/E	S/N	F/T	J/P
Bharadwaj, et al. (2018)	SVM, MLP and Naïve Bayes	MBTI_Kaggle	NB	Accuracy	77%	86.2%	77.9%	62.3%
				Recall				
				Precession				
			SVM	Accuracy	84.9%	88.4%	87.0%	78.8%
				Recall				
				Precession				
			MLP	Accuracy	77.0%	86.3%	54.1%	61.8%
				Recall				
				Precession				
Gjurković et al. (2018)	SVM, MLP and Logistic Regression	MBTI9k	SVM	Accuracy				
				F1-Score	79.6%	75.6	64.8	72.6
				Precession				
			LR	Accuracy				
				F1-Score	81.6	77.0	67.2	74.8
				Precession				
			MLP	Accuracy				
				F1-Score	82.8	79.2	64.8	72.6
				Precession				
Proposed (our work)	XGBoost	MBTI_Kaggle						
			XGBoost	Accuracy	99.37	99.92	94.55	95.53
				Recall	97.16	100	89.96	92.66
				Precession	100	99.50	100	100
				F1-Score	98.56	99.75	94.72	96.19

V. CONCLUSION AND FUTURE WORK

The central theme of this study is the application of different machine learning techniques on the benchmark, MBTI personality dataset namely mbti_kaggle to classify the text into different personality traits such as Introversion-Extroversion(I-E), iNtuition-Sensing(N-S), Feeling-Thinking(F-T) and Judging-Perceiving(J-P).

The Mayers-Briggs Type Indicator (MBTI) model is used for text classification and personality traits recognition [4]. After applying class balancing techniques on the imbalanced classes, different machine learning classifiers, namely, KNN, Decision Tree, Random Forest, MLP, Logistic Regression (LR), SVM, XGBoost, MNB and Stochastic Gradient Descent (SGD) are experimented to identify the personality traits. Evaluation metrics, such as accuracy, precision, recall and F-score, are used to analyze and examine the overall efficiency of the predictive model. The obtained results show that score achieved by all classifiers across all personality traits is good enough, however, the performance of XGBoost classifier is outstanding. We got more than 99% precision and accuracy forI/E and S/N traits and obtained all about 95% accuracy for T/F and J/P dimensions. However, KNN classifier resulted in overall lower performance.

A. Constraints or Limitations

1) MBTI model is examined for personality traits classification, however, others personality models such as Big Five Factor (BFF) and DiSC personality Assessment models, are not experimented and investigated.

2) The textual data used in the proposed work for personality assessment is comprised of only English language, and the contents of other languages are not experimented.

3) Simple over-sampling and under sampling techniques are used to balance and level the skewness of dataset.

4) The dataset comes from only one platform namely personalitycafe forum, which may lead to biased results.

5) All the experiments conducted in this proposed work are based on the classical or traditional machine learning algorithms.

6) The textual contents which are classified for personality traits identification belong to only one site Twitter, however other social networking sites are ignored.

7) Only textual data is analysed and investigated for user's personality traits recognition in his proposed work.

8) Less weightage is given to feature extraction in classification of text, only TF-IDF technique is utilized.

B. Future Proposal

1) The predictive performance of MBTI personality model needs to be compared with the Big Five Factor (BFF) model for better assessment of the traits.

2) Multilingual textual content, especially Urdu and Pashto language textual data can be examined for personality classification.

3) SMOTE (Synthetic Minority Over-sampling Technique) can be utilized as class balancing method for more robust and reliable performance.

4) Labelled data may need to be collected from other platforms like “Reddit” using multiple benchmark datasets.

5) More experiments on personality recognition may be conducted using Deep learning algorithms.

6) Other social networking sites like FACEBOOK posts and comments are required to be examined for automated personality traits inference.

7) Data available in the format of images and videos on social networking sites can be experimented for the task of personality traits identification.

8) More advanced features selection approaches are required to be exploited for enhancement of the proposed work.

References

[1] N. Majumder, S. Poria, A. Gelbukh and E. Cambria, "Deep Learning-Based Document Modeling for Personality Detection from Text," in IEEE Intelligent Systems, vol. 32, no. 2, pp. 74-79, Mar.-Apr. 2017.

[2] D. Xue et al., "Personality Recognition on Social Media With Label Distribution Learning," in IEEE Access, vol. 5, pp. 13478-13488, 2017.

[3] L. R. Goldberg, L. R. ,”An alternative" description of personality": the big-five factor structure,” Journal of personality and social psychology, vol. 59, no. 6, p.1216, 1990

[4] I. B. Myers, “The Myers-Briggs Type Indicator: Manual” ,1962

[5] D. Shaffer, M. Schwab-Stone and P. Fisher, “Preparation, field testing, interrater reliability and acceptability of the DIS-C,” J Am Acad Child Adolesc Psychiatry, vol. 32, pp. 643-648, 1993.

[6] S. Bharadwaj, S. Sridhar, R. Choudhary and R. Srinath, "Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification Approach," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, 2018, pp. 1076-1082.

[7] M. Gjurković and J. Šnajder, “Reddit: A Gold Mine for Personality Prediction,” In Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media , pp. 87-97, 2018.

[8] B. Plank, and D. Hovy, “Personality traits on twitter—or—how to get 1,500 personality tests in a week.” In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 92-98, 2015.

[9] O. Loyola-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa and M. García-Borroto, “Study of the impact of resampling methods for contrast pattern-based classifiers in imbalanced databases,” Neurocomputing, 175, pp. 935-947, 2016.

[10] A. More, “Survey of resampling techniques for improving classification performance in unbalanced datasets,” 2016, arXiv preprint arXiv:1608.06048

[11] P. Kaur and A. Gosain, “Comparing the Behavior of Oversampling and Undersampling Approach of Class Imbalance Learning by Combining Class Imbalance Problem with Noise,” In ICT Based Innovations , pp. 23-30, Springer, Singapore, 2018.

[12] I. Cantador, I. Fernández-Tobías and A. Bellogín, “Relating personality types with user preferences in multiple entertainment domains,” In CEUR workshop proceedings, ShlomoBerkovsky, 2013.

[13] B. Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," 2015 International Conference on Data and Software Engineering (ICoDSE), Yogyakarta, 2015, pp. 170-174.

[14] V. Ong et al., "Personality prediction based on Twitter information in Bahasa Indonesia," 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), Prague, 2017, pp. 367-372.

[15] F. Alam, E. A. Stepanov and G. Riccardi, “Personality traits recognition on social network-facebook,” WCPR (ICWSM-13), Cambridge, MA, USA, 2013.

[16] K. Buraya, A. Farseev, A. Filchenkov and T. S. Chua, “Towards User Personality Profiling from Multiple Social Networks,” In AAAI, pp. 4909-4910, 2017.

[17] N. R. Ngatirin, Z. Zainol and T. L. C. Yoong, "A comparative study of different classifiers for automatic personality prediction," 2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Batu Ferringhi, 2016, pp. 435-440.

[18] S. Chaudhary, R. Sing, S. T. Hasan and I. Kaur, “A comparative Study of Different Classifiers for Myers-Brigg Personality Prediction Model,” IRJET, vol.05, pp.1410-1413, 2018.

[19] V. Ong, A. D. Rahmanto, Williem and D. Suhartono, “Exploring Personality Prediction from Text on Social Media: A Literature Review,” INTERNETWORKING INDONESIA, vol. 9, no. 1, pp. 65-70, 2017a.

[20] J. Golbeck, C. Robles, M. Edmondson and K. Turner, "Predicting Personality from Twitter," 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Boston, MA, 2011, pp. 149-156.

[21] D. Quercia, M. Kosinski, D. Stillwell and J. Crowcroft, "Our Twitter Profiles, Our Selves: Predicting Personality with Twitter," 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Boston, MA, 2011, pp. 180-185.

[22] B., Verhoeven, W. Daelemans and B. Plank, “Twisty: a multilingual twitter stylometry corpus for gender and personality profiling,” In Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016)/Calzolari, Nicoletta [edit.]; et al. pp. 1-6, 2016.

[23] F. Celli, “Mining user personality in twitter, “ Language, Interaction and Computation CLIC, 2011.

[24] X. Sun, B. Liu, Q. Meng, J. Cao, J. Luo and H. Yin, “Group-level personality detection based on text generated networks,” World Wide Web, pp. 1-20, 2019.

[25] F. Celli and L. Rossi, “The role of emotional stability in Twitter conversations,” In Proceedings of the workshop on semantic analysis in social media, Association for Computational Linguistics, pp. 10-17, 2012.

[26] S. Chishti, X. Li and A. Sarrafzadeh, “Identify Website Personality by Using Unsupervised Learning Based on Quantitative Website Elements, “ In International Conference on Neural Information Processing, Springer, Cham. pp. 522-530, 2015.

[27] F. Celli, “Unsupervised personality recognition for social network sites,” In Proc. of Sixth International Conference on Digital Society, 2012.

[28] P. H. Arnoux, A. Xu, N. Boyette, J. Mahmud, R. Akkiraju and V. Sinha, “25 Tweets to Know You: A New Model to Predict Personality with Social Media,” 2017, arXiv preprint arXiv:1704.05513.

[29] M. Arroju, A. Hassan, and G. Farnadi, “Age, gender and personality recognition using tweets in a multilingual setting,” In 6th Conference and Labs of the Evaluation Forum (CLEF 2015): Experimental IR meets multilinguality, multimodality, and interaction, pp. 23-31, 2015.

[30] L. C. Lukito, A. Erwin, J. Purnama and W. Danoekoesoemo, "Social media user personality classification using computational linguistic," 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, 2016, pp. 1-6.

[31] N. Alsadhan and D. Skillicorn, "Estimating Personality from Social Media Posts," 2017 IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, 2017, pp. 350-356.

[32] R. K. Hernandez and L. Scott, "Predicting Myers-Briggs type indicator with text," In *31st Conference on Neural Information Processing Systems* (*NIPS 2017)*, 2017.

[33] B. Cui and C. Qi, "Survey Analysis of Machine Learning Methods for Natural Language Processing for MBTI Personality Type Prediction".

[34] D. Xue, L. Wu, Z. Hong, S. Guo, L. Gao et al, "Deep learning-based personality recognition from text posts of online social networks," *Applied Intelligence*, vol. 48, no. 11, pp. 4232-4246, 2018.

[35] Y. Yan, Y. Liu, M. Shyu and M. Chen, "Utilizing concept correlations for effective imbalanced data classification," Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014), Redwood City, CA, 2014, pp. 561-568.

[36] S. Rezaei and X. Liu, "Deep Learning for Encrypted Traffic Classification: An Overview," in IEEE Communications Magazine, vol. 57, no. 5, pp. 76-81, May 2019.

[37] M. Z. Asghar, A. Khan, F. Khan and F. M. Kundi, "RIFT: A Rule Induction Framework for Twitter Sentiment Analysis," Arabian Journal for Science and Engineering, vol. 43, no. 2, pp.857-877, 2018.

[38] A. Tripathy, A. Agrawal and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," Expert Systems with Applications, 57, pp. 117-126, 2016.

[39] L. H. Patil and M. Atique, "A novel approach for feature selection method TF-IDF in document clustering," 2013 3rd IEEE International Advance Computing Conference (IACC), Ghaziabad, 2013, pp. 858-862.

[40] M. C. Komisin and C. I. Guinn, "Identifying personality types using document classification methods," In Twenty-Fifth International FLAIRS Conference, 2012.

[41] D. Nielsen, "Tree Boosting With XGBoost-Why Does XGBoost Win Every Machine Learning Competition? (Master's thesis, NTNU)," 2016.

[42] M. M. Tadesse, H. Lin, B. Xu and L. Yang, "Personality Predictions Based on User Behavior on the Facebook Social Media Platform," in IEEE Access, vol. 6, pp. 61959-61969, 2018.

Control BLDC Motor Speed using PID Controller

Md Mahmud[1], S. M. A. Motakabber[2], A. H. M. Zahirul Alam[3], Anis Nurashikin Nordin[4]
Department of Electrical and Computer Engineering
International Islamic University Malaysia
Kuala Lumpur, Malaysia

***Abstract*—At present, green technology is a major concern in every country around the world and electricity is a clean energy which encourages the acquisition of this technology. The main applications of electricity are made through the use of electric motors. Electric power is converted to mechanical energy using a motor, that is to say, the major applications of electrical energy are accomplished through electric motors. Brushless direct current (BLDC) motors have become very attractive in many applications due to its low maintenance costs and compact structure. The BLDC motors can be substituted to make the industries more dynamic. To get better performance BLDC motor requires control drive facilitating to control its speed and torque. This paper describes the design of the BLDC motor control system using in using MATLAB/SIMULINK software for Proportional Integral Derivative (PID) algorithm that can more effectively improve the speed control of these types of motors. The purpose of the paper is to provide an overview about the functionality and design of the PID controller. Finally, the study undergoes some well-functioning tests that will support that the PID regulator is far more applicable, better operational, and effective in achieving satisfactory control performance compared to other controllers.**

Keywords*—PID controller; green technology; fuzzy logic control; speed control; BLDC motor

I. Introduction

The present era is the era of the industrial revolution, which began with the invention of motor. Various types of motors have been developed over time, but these motors are generally classified into two main categories, namely, AC motor and DC motor. There exists a set of DC motors that can be used on different devices. However, generally two types of DC motors are set up in industrial applications. In the first type, the magnetic flux is generated by the current through the field coil of static pole structure and in the second type, permanent magnet supplies the required air gap flux [1]. A BLDC motor is a special type of DC motor that does not apply a brush for transport, instead an electronic process system is used for this purpose. The BLDC motor is usually a synchronous motor composed of a trapezoidal back EMF waveform and a permanent magnet. The current trend shows that high-performance BLDC motor technologies are widely used for global industrial applications and variable speed drives in electric vehicles [2]. In fact, these types of motors depend on its control circuit. In fact, these types of motors rely on its control circuit and still developing a high performance circuit is a challenging task for researchers. A basic control system is shown in Fig. 1 for the BLDC motor.

The structure of the BLDC motor tuning control project selection, modelling simulation and so on. The design structure of a BLDC motor is a complex task and depends on many issues such as project selection, modeling, simulation, etc. In terms of the rapidity framework of the BLDC motor, a host of modern control solutions have been proposed [3].

The key features of a conventional PID controller algorithm are it is easily adjustable, steady operation and its simple design, which making it widely used for controlling system. For practical reason, common speed control structure is applied in the PID controller. The mathematical model and speed control of the BLDC motor have been proposed and validated using fuzzy logic and PID controller [4]. Most of the cases a different finding is seen in terms of practical utility experiences where the volatility of well-structured prototype, different units of nonlinear, low variability have been at work. For tuning a PID controller parameters are not that simple, hence, getting the optimal position under the examined circumstances is challenging [5]. This study proposes a PID controller through modifying some changes thereto which, may increase the regulation speed of BLDC motor. In this case parameters can be tuned at the actual moment under PID controller operation. In the sake of better functioning of the PID controller scheme requires input and membership function enhancement [6]. At the same time, a set of values are applied for the PID controller's constant coefficients, K_p, K_i and K_d. By employing these values, the proposed modified controller would be restructured to any adjusting dimension.

The purpose of this study is to show the dynamic response to the rapid tuning results of the proposed modified PID controller; which can help to control the speed of the motor and to maintain constant speed during load changes. Thus, the PID regulator can increase the overall performance of the BLDC motor. The simulation results showed that the functions of the PID controller could be provided with a better control performance [7].

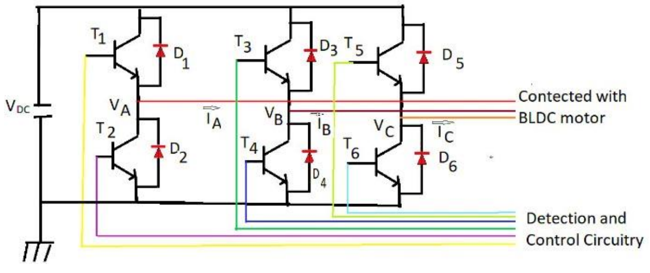


Fig 1. Basic Circuit Diagram for the BLDC Motor Control System.

II. BLDC Motor and Speed Control System

A. Speed Control System

A controller circuit is essential to operate and control the speed of a BLDC motor. There are many types of speed control system developed for controllers but the speed controllers have to modernize with the ages. However, they are generally classified as closed loop and open loop control systems, respectively. Closed loop techniques are used for high accuracy control system. Fig. 2 shows a BLDC motor speed controller block diagram using two closed loop systems. In this case, the internal loop is used for tuning and sense the power supply polarity and the external loop is used to control the speed. The motor speed controller helps to adjust the voltage of the DC bus. To control the system, DC supply is required and its value depends on the motor speed (rpm) and its capacity. This system also requires a controller, in which case a PID controller is used that ultimately controls the inverter output voltage. A sensor is an integral part of a closed loop controller for controlling the speed of a motor. The primary function of the sensor is to convert the physical position and condition of the motor shaft into an equivalent electrical signal for the controller circuit. Typically, BLDC motor requires an AC-like voltage-waveform for its operation, so inverter circuit is used to convert the DC power supply voltage into an equivalent AC supply voltage [8], [9] for proper function.

B. The Back Electro Motive Fource (BEMF)

Typically, a 3-phase BLDC motor uses six electronic switches (power transistors) to produce 3-phase voltage simultaneously to a full-bridge configuration power converter. The transistors have a rotor position, which will be defined as the switching sequence. Most of the cases motor starter is monitoring by using three hall sensor devices. The hall sensors provide the information to the decoder block for producing the sign of reference current signal vector to the back electromotive force (BEMF). To operate the motor in the opposite direction, the current is changed in reverse direction or the switching order of the controller is changed.

The MATLAB simulation block diagram for generating the back EMF of the decoder is shown in Fig. 3, and Table I shows the decoder sequences of the proposed 3-phase PID controller for the BLDC motor to rotate in the clockwise direction.

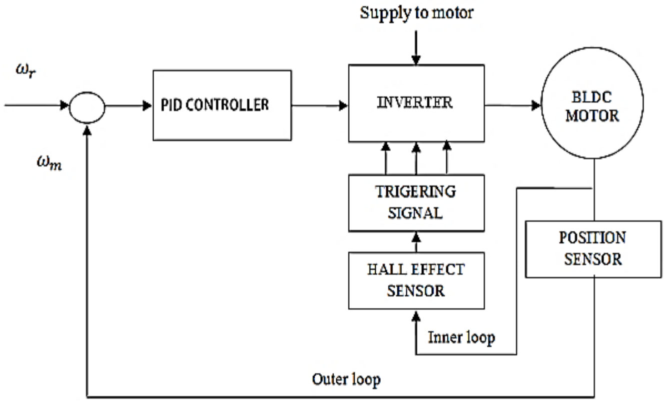


Fig 2. Block Diagram of BLDC Motor Speed Control.

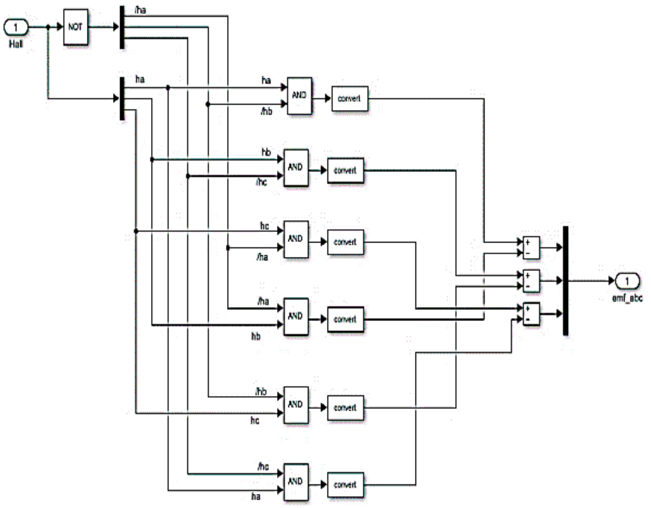


Fig 3. Back EMF of Decoder for MATLAB Drive.

TABLE I. True Table for Decoder

Table I.	True Table For Decoder					
	ha	*hb*	*hc*	*emf_a*	*emf_b*	*emf_c*
	0	0	0	0	0	0
	0	0	1	0	-1	+1
	0	1	0	-1	+1	0
	0	1	1	-1	0	+1
	1	0	0	+1	0	-1
	1	0	1	+1	-1	0
	1	1	0	0	+1	-1
	1	1	1	0	0	0

Similarly, Fig. 4 shows the functional block diagram of the inverter switching for MATLAB simulation, and Table II shows the decoder sequences of the proposed 3-phase PID controller for the BLDC motor to rotate in the counterclockwise motion.

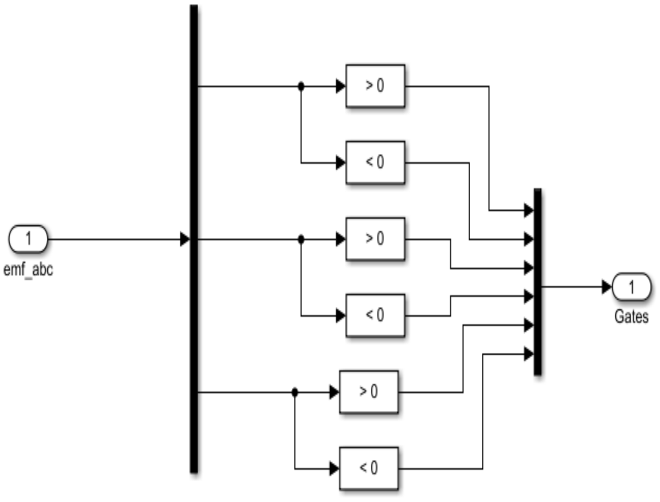


Fig 4. Inverter Switching for MATLAB Drive.

TABLE II. TRUE TABLE FOR INVERTER SWITCHING

Table II.	**True Table For Inverter Switching**								
	emf_a	***emf_b***	***emf_c***	***Q1***	***Q2***	***Q3***	***Q4***	***Q5***	***Q6***
	0	0	0	0	0	0	0	0	0
	0	-1	+1	0	0	0	1	1	0
	-1	+1	0	0	1	1	0	0	0
	-1	0	+1	0	1	0	0	1	0
	+1	0	-1	1	0	0	0	0	1
	+1	-1	0	1	0	0	1	0	0
	0	+1	-1	0	0	1	0	0	1
	0	0	0	0	0	0	0	0	0

III. PROPOSED PID CONTROLLER MODELING

For getting better performance of DC motors it is essential to use a controller circuit. For this purpose, a variety of controller circuits and algorithms are used. However, among them PID controller is the most suitable controller circuit for BLDC motor. The PID controller is mainly composed of three block of circuits and they are proportional, integral and derivative blocks. Each block of circuit is used to perform different mathematical operations as their name mentioned. The complete MATLAB design of the proposed controller for 3-phase brushless DC motor is shown in Fig. 5. The diagram clearly shows how the reference source, PID controller, driver circuit, sensors, converter circuit, inverter circuit, display scope and motor are interconnected.

The foundational frequency transferring performance G(s) of the PID controller can be represented by (1) and (2),

$$G(s) = K_p + K_i/s + K_d s \quad (1)$$

$$G(s) = (Kd\ s2 + Kp\ s + Ki)/s \quad (2)$$

Where, K_p = proportional gain coefficient, K_i = integral gain coefficient, K_d = derivative gain coefficient and s is the complex frequency.

The time derivative output U(t) of the controller for control of the plant is equal to K_p times the magnitude of error pulse K_d times the derivative of time function error signal e(t) and K_i times the integral can be represented by (3).

$$U(t) = Kp\ e(t) + Ki \int e(t)dt + Kd\ de(t)/dt \quad (3)$$

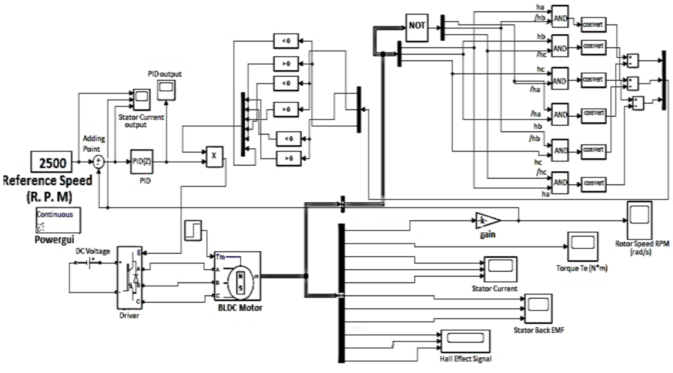


Fig 5. Complete MATLAB Design of Controller for BLDC Motor.

Its applications are wide because of its ease and outstanding performance, in many cases its efficiency is more than 95%. Typically a closed-loop PID controller is used for industry application. The four key features are most interested in the response to the closed-loop step, they are, settling time, overshoot, steady-state error and response time.

Table III shows the values of the PID controller parameters used for this design.

TABLE III. VALUES OF THE PID CONTROLLER PARAMETERS

Table III	**Values Of The PID Controller Parameters**			
	Method	K_p	K_i	K_d
	PID	100	0.5	500

IV. RESULT AND DISCUSSION

The performance of the proposed PID controller for brushless DC motor at 2500 rpm is shown in Fig. 6. In this figure, X and Y axis represented the time in second (sec) and the speed (rpm) of the BLDC motor at no-load condition respectively. From the figure, it is seen that the settling time of the controller is about 0.018 sec with a negligible amount of overshoot and undershoot. After 0.018 sec the motor runs at a constant speed of the preset value 2500 rpm.

Fig. 7 shows the output torque response performance of the BLDC motor at no-load condition. In this figure, X and Y axis represented the time in second (sec) and electromagnetic torque value in Newton-meter (Nm) of the BLDC motor at no-load condition, respectively. From the figure, it is observed that the motor electromagnetic torque (Nm) is fixed after about 0.030 sec.

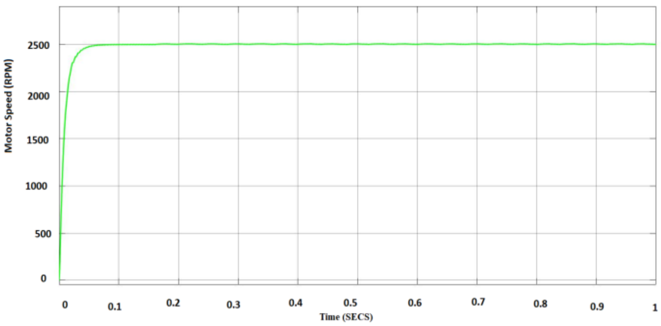


Fig 6. The no-load performance of the BDC motor using.

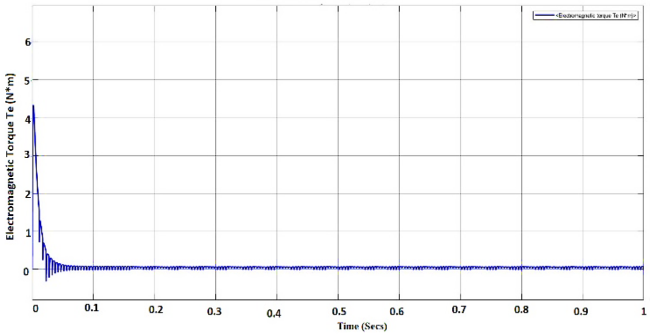


Fig 7. Time Versus Electromagnetic Torque Response of the BLDC Motor.

Fig. 8 shows the stator current of the BLDC motor at no-load condition. In this figure X-axis represented the time in second (secs) and Y-axis the motor stator current in Ampere (A) respectively. In this figure the 3-phase stator currents are illustrated by the green, pink and yellow color lines respectively. It is also observed that the stators current are fixed after 0.030 (secs) which is the same as electromagnetic torque fixing time of the result as shown in Fig. 6.

Fig. 9 shows the back electromotive force (emf) of the BLDC motor at no-load condition. In this figure, X and Y axis represented the time in second (secs) and back emf value in Volt (V) of the BLDC motor at no-load condition respectively. The 3-phase back emf voltages of the BLDC motor are illustrated by the green, pink and yellow color lines in this figure, respectively. It is clear from the figure that the 3-phase back emf voltages are fixed ±24V after 0.030 (secs), which justifies the result same as the results obtained in Fig. 6 and Fig. 8.

Fig. 10 shows the 3-phase signals generated from Hall-sensor. Here the green, pink and yellow color lines are represented the individual phase signal generated by the sensor.

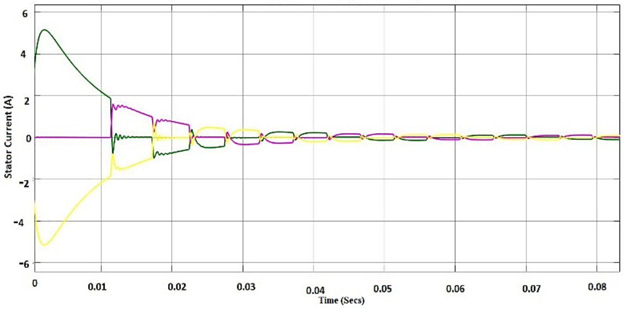


Fig 8. Phase Stator Current of the BLDC Motor.

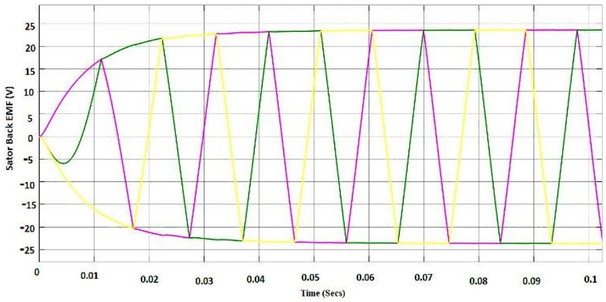


Fig 9. BLDC Motor 3-Phase Back Electromotive Force.

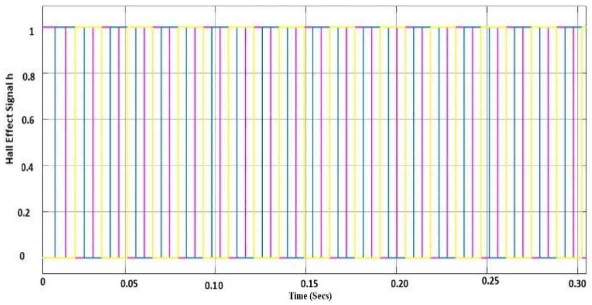


Fig 10. BLDC Motor Hall Effect Signal.

Fig. 11 shows the output performance of the PID controller. The X-axis and Y-axis represented the second (secs) and reference signal value in RPM of the PID controller, respectively. From the diagram, it is seen that the PID controller output under shoot and reach minimum value about 0.03 sec, then it reach at a stable condition after about 0.03 sec.

Fig. 12 shows the performance comparison of the PI, PID and Fuzzy logic controller. In this figure X-axis and Y-axis represented the time in second (secs) and electromagnetic torque value in Newton-meter (Nm) of the BLDC motor at no-load condition, respectively. Here, the red, yellow and green color lines are illustrated the PI, PID and Fuzzy logic controllers output performance in some respects. It is noteworthy that the performance of the PID controller is better than the other two controllers. The BLDC motor predetermined reference speed of 2500 rpm has been chosen in this study.

From Fig. 6 and Fig. 12 it is observed that the PID controller settlement time about 18 milliseconds (msecs). The overshoots and undershoots of this controller are 0.4% and 1.9%, respectively, which is within the tolerable range of a BLDC motor for suitable operation. The PID controller slew rate about 92.27 (msecs) and per shoot 2.5%.

On the other hand, PI controller rise time is better than the PID controller but, its slow rate is much higher as 621.35 (msecs). In the other aspects of the PI controller, the pre-shoot 0.66%, overshoot 32.67%, undershoot 1.68% and settling time 15.20 msec. The pre-shoot, overshoot and undershoot are basically a high frequency nose and can be minimized by using filter [10].

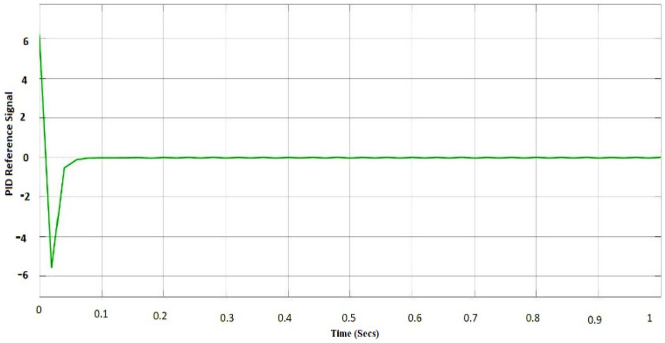


Fig 11. PID Controller Output Performance with Time.

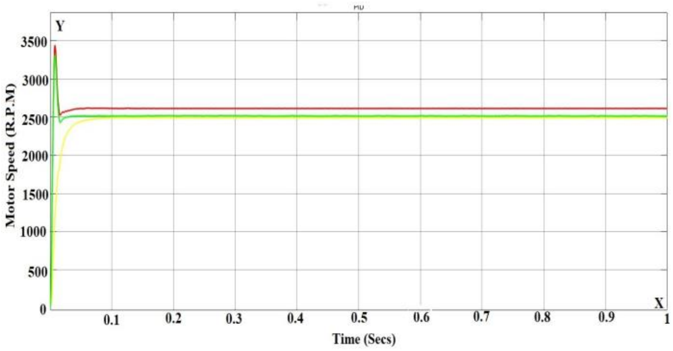


Fig 12. Performance Comparison among PI, PID and Fuzzy Logic Controllers.

It shows that the output of the Fuzzy logic controller, the pre-shoot 0.67%, slew rate 598.15 (msecs), overshoot 30.92%, undershoot 3.2% and settling time 9.2 (msecs).

From comparative performance analysis it is understandable that the PID controller will provide the best results for BLDC motor control.

V. Conclusion

A three-phase BLDC motor controller has been successfully designed based on PID controller scheme and compared its performance with PI and Fuzzy logic controllers. From the results, it is observed that the PID controller provides the best performance compared to the other two controllers, PI and Fuzzy logic. The design has been validated by MATLAB simulation.

Acknowledgment

This research has been supported by the Malaysian Ministry of Education through the Fundamental Research Grant Scheme under the project ID: FRGS19-054-0662.

References

[1] H. Salim Hameed, "Brushless DC Motor Controller Design Using Matlab Applications," 3rd Scientific Conference of Engineering Science (ISCES), 19 April 2018.

[2] M. Rajkumar, G. Ranjhitha, G. Pradeep and M. F. Kumar, "Fuzzy-based Speed Control of Brushless DC Motor feed electric vehicle," IJISSET, 2017, vol. 3(3).

[3] K. Swapnil, J. Anjali, A. Mohan and D. Shantanu, "Modeling and control of a permanent-magnet brushless DC motor drive using a fractional-order proportional-integral-derivative controller," Turkish Journal of Electrical Engineering and Computer Sciences, 2017, 25(5), pp. 4223-4241.

[4] M. Valan Rajkumar, G. Ranjhitha, M. Pradeep, M. Fasil PK and R. Sathish Kumar, "Fuzzy based Speed Control of Brushless DC Motor fed Electric Vehicle," International Journal of Innovative Studies in Sciences and Engineering Technology (IJISSET), 2017, vol. 3(3).

[5] H. Manal and Jasim, "Tuning of a PID Controller by Bacterial Foraging Algorithm for Position Control of DC Servo Motor," Iraqi Academic Scientific Journals, 2018, vol. 36.

[6] M. Singirala, D. Krishna and T. Anil Kumar, "Improving Performance Parameters of PMBLDC Motor using Fuzzy Sliding Mode Controller," International Journal of Recent Technology and Engineering (IJRTE), 2019, vol. 8(4).

[7] M. Rafay Khan, A. Ahmed Khan and U. Ghazali, "Speed Control of DC Motor under Varying Load Using PID Controller," International Journal of Engineering (IJE), 2015, vol. 9(3).

[8] T. Rahman, S. M. A. Motakabber, M. I. Ibrahimy and A. H. M. Zahirul Alam, "PLL-Based 3φ Inverter Circuit for Microgrid System Operated by Electrostatic Generator," IIUM Engineering Journal, 2019, vol. 20(1), pp. 177-193.

[9] T. Rahman, S. M. A. Motakabber, M. I. Ibrahimy and Aliza 'Aini, "A PWM Controller of a Full Bridge Single-phase Synchronous Inverter for Micro-grid System," Journal of Physics: Conference Series, 2017, pp.1-13.

[10] S. M. A. Motakabber, M. A. Mohd Ali and N. Amin, "Computer Aided Design of an Active Notch Filter for HF Band RFID," FREQUENZ, 2010, vol. 64(1-2), pp. 23-25.

Recurrent Neural Networks for Meteorological Time Series Imputation

Anibal Flores[1]
Grupo de Investigación en Ciencia de Datos, Universidad Nacional de Moquegua, Moquegua, Perú

Hugo Tito[2]
E.P. Ingeniería de Sistemas e Informática, Universidad Nacional de Moquegua, Moquegua, Perú

Deymor Centty[3]
E.P. Ingeniería Ambiental
Universidad Nacional de Moquegua
Moquegua, Perú

***Abstract*—The aim of the work presented in this paper is to analyze the effectiveness of recurrent neural networks in imputation processes of meteorological time series, for this six different models based on recurrent neural networks such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are implemented and it is experimented with hourly meteorological time series such as temperature, wind direction and wind velocity. The implemented models have architectures of 2, 3 and 4 sequential layers and their results are compared with each other, as well as with other imputation techniques for univariate time series mainly based on moving averages. The results show that for temperature time series on average the recurrent neural network achieve better results than the imputation techniques based on moving averages; in the case of wind direction time series, on average only one model based on RNN manages to exceed the models based on moving averages; and finally, for wind velocity time series on average, no RNN-based model manages to exceed the results achieved by moving averages-based models.**

***Keywords*—Recurrent neural network; long short-term memory; gated recurrent unit; univariate time series imputation**

I. Introduction

The imputation of time series is a very important activity within the stage of homogenization of data, it is typical of the processing of meteorological time series. This will allow a subsequent time series to be used in forecasting processes.

There are many reasons why NA values are found: values may not have been measured, values may be measured but lost or values may be measured but erroneously [1]. Missing values can cause problems, since complete data is usually needed for proper processing and analysis.

It's very known that the accuracy of the imputation techniques will allow better results in forecasting or prediction processes [2]. Thus, a good selection of the imputation technique presented a certain problem is very important.

There is not a very large number of imputation techniques for univariate time series, among them can be mentioned those based on moving averages such as: Simple Moving Average (SMA) [3], Linear Weighted Moving Average (LWMA) [3], Exponential Weighted Moving Average (EWMA) [3], Autoregressive Integrated Moving Average (ARIMA) [4] among others.

Nowadays, recurrent neural networks (RNN) such as Long Short-Term Memory (LSTM) [5] and Gated Recurrent Unit GRU) [6] have become the most commonly used in prediction or forecasting models today for the accuracy of the results they offer in different fields such as machine translation, robot control, speech recognition, time series prediction among others. However, despite the benefits described, in the state of the art it is very difficult to find works that use recurrent neural networks for univariate time series imputation, which was one of the main motivations for the realization of the present study.

Thus, this paper presents the results of the implementation of six different models for hourly time series imputation based on recurrent neural networks. The analyzed time series correspond to temperature, wind direction and wind velocity and they were obtained from the Moquegua[1] meteorological station of SENAMHI in southern Peru. The gap-sizes analyzed correspond to short-gaps (1 to 2 NAs), medium-gaps (3 to 10 NAs) and large-gaps (11 to 30 NAs) [7]. Fig. 1 shows a graphical view of the 3-time series for 24 hours.

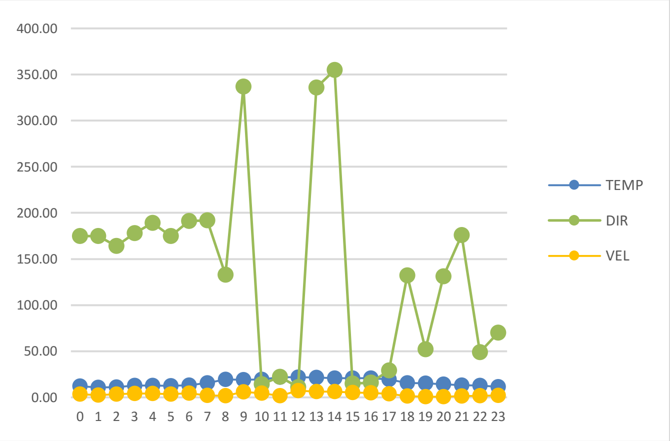


Fig. 1. Hourly Temperature, wind Direction and wind Velocity.

The content of the paper has been organized as follows: In the second section the related work is briefly described as proposed in this study. In the third section, the theoretical concepts and bases that will allow a better understanding of the content of the paper are described. In the fourth section, the models based on recurrent neural networks implemented in this

[1] SENAMHI Lat. 17°10'9" Lon. 70°55'54" Alt. 1450 masl. https://www.senamhi.gob.pe/?&p=estaciones

study are described and detailed. In the fifth section, the results achieved by the six different models in the time series are explained in detail. In the sixth section, the results achieved by the proposed models are compared and discussed with other models and techniques of the state of the art. In the seventh section, the arrived conclusions are explained according to study results. Finally, it describes the future work that can be done to improve the achieved results.

II. Related Work

This section shows a brief review of the works related to this study which are described below:

The first methods of imputation consisted of the use of parameters such as mean, median or mode [8], due to its simplicity there was a risk of inserting bias into the time series.

Another technique used later than the first was to use the last data observed before the missing one. This was called Last Observation Carried Forward (LOCF).

We also have the Hot-Deck [9] technique that consisted of randomly using existing data to replace the Not Available (NA) value.

Another group of techniques widely used are those based on moving averages including Simple Moving Average (SMA) [3], Linear Weighted Moving Average (LWMA) [3], Exponential Weighted Moving Average (EWMA) [3] which basically consisted of using the average of the data around the missing data assigning a weight according to its proximity to the NA value. This set of techniques are implemented in the present study to compare the results achieved by the imputation models based on recurrent neural networks.

An improved technique based on moving averages is what is known as Autoregressive Integrated Moving Averages (ARIMA) [4], which is a statistical technique that works with variations and regressions in a series of time to find patterns that will later serve to make predictions. This work also compares the results of ARIMA with those achieved by the imputation models based on recurrence neural networks.

Another technique used for imputation of time series is known as Local Average of Nearest Neighbors (LANN) [2], this technique is quite simple and consists only of using the prior and next values around an NA value, producing very good results at the level or better than those based on moving averages.

Two new imputation techniques inspired by Case Based Reasoning [10] are CBRi [11] and CBRm [8] which, like LANN, use only the prior and next values of an NA value, completing the missing values from the average of the historical data similar to the prior and next. The difference between the two is that CBRi is designed for short-gaps and CBRm for medium-gaps.

Another new technique is known as Average of Historical Vectors (AHV) [12] that uses only values similar to the prior value of the NA value to calculate the missing data. This technique is complemented by an adjustment algorithm (iNN) [12] and a smoothing algorithm (LANNf) [12].

III. Background

A. Time Series Imputation

The time series imputation refers to the process of calculating and completing the missing data or Not Available (NA) values in a series of time. For this it is very important to determine how the NA values originated, so they can be Missing Completely at Random (MCAR), Missing at Random (MAR) or Not Missing at Random (NMAR) [1]. It is also very important to determine the characteristics of the time series, so it can be very useful some characteristics such as: trend, seasonal or non-seasonal cycles, pulses, etc.

B. Recurrent Neural Networks (RNN)

An RNN is a type of neural network [13] that allows modeling different kind of problems such as time series for prediction.

The architecture of this neural network is very similar to the architecture of a Multilayer Perceptron (MLP) with the difference that an MLP allows connections between hidden units associated with a time delay. These connections allow the RNN to retain and remember information from the past [14], in this way it can find temporary correlations between facts that can be very separated in time. Fig. 2 shows the unfolded structure of an RNN.

Training an RNN is very difficult to implement [13] due to the vanishing and exploding gradients, This led to the implementation of a special type of RNN that is known as LSTM (Long Short-Term Memory) and that solves the above problems.

C. Long Short-Term Memory (LSTM)

As mentioned above, the LSTM networks were created to solve the problem of the vanishing and exploding gradients of the first recurrent neural networks. The LSTM networks work with special hidden units, whose objective is to remember input data for a long time [3], so LSTM networks are better than conventional RNN [5]. LSTM networks have several layers for each time step. Fig. 3 shows the LSTM architecture.

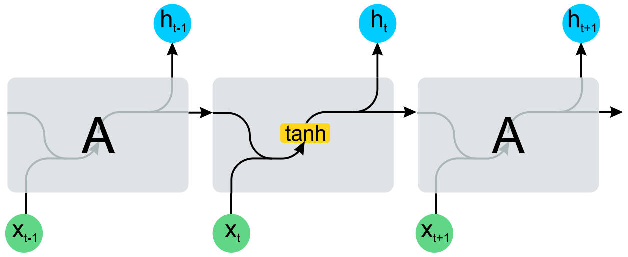


Fig. 2. Architecture of Recurrent Neural Network.

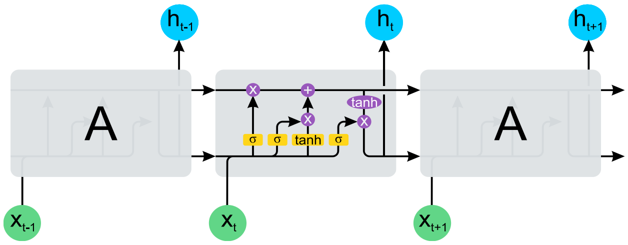


Fig. 3. Architecture of LSTM Network.

D. Gated Recurrent Unit (GRU)

GRUs are an activation mechanism in RNNs and were introduced by K. Cho et al. [6] in 2014. GRUs are a variation of LSTM networks, since both have a very similar architecture. However, unlike LSTM networks, GRUs have fewer parameters, since they lack an output gate. In many studies, LSTM networks have proven to be stronger than GRUs, since they can easily perform unlimited counting, while GRUs do not, so GRUs do not learn certain languages that LSTM can do [15]. Fig. 4 shows a very common architecture of GRU.

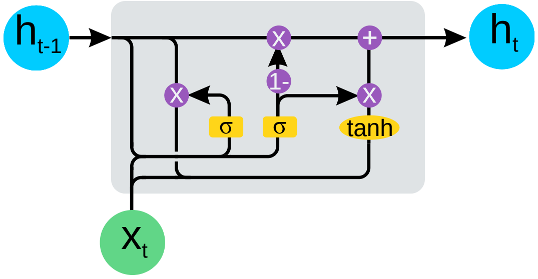


Fig. 4. Architecture of GRU.

According to Fig. 3 the following equations can be obtained and some parameters are described:

$$Z_{t=}\sigma_g(W_z x_t + U_z h_{t-1} + b_z) \quad (1)$$

$$r_{t=}\sigma_g(W_r x_t + U_r h_{t-1} + b_r) \quad (2)$$

$$h_{t=}(1 - z_t) \text{ o } h_{t-1} + z_t \text{ o } \sigma_h(W_h x_t + U_h(r_t \text{ o } h_{t-1}) + b_h \quad (3)$$

Where:

x_t : input vector

h_t : output vector

z_t : updated gate vector

r_t : reset gate vector

W,U and *b* : matrix parameters and vector

σ_g : sigmoid function

σ_h : hyperbolic tangent

IV. Models for Experimentation

In the present study, six models based on recurrent neural networks were implemented, which are described below:

As can be seen in Table I of the six models implemented, three correspond to LSTM and three to GRU, the process followed to implement each of them is described below.

A. Time Series Selection

The hourly time series chosen for experimentation corresponds to temperatures, wind direction and wind velocity obtained from the SENAMHI repository. The data used for the training stage corresponds to 6000 hours from 2019-05-20 00:00:00 to 2020-01-24 23:00:00. The same period was used for all three-time series.

Likewise, the data between 2020-01-25 00:00:00 and 2020-01-31 23:00:00 was chosen as testing data (168 hours).

TABLE. I. RNN Models

Model	Name	RNN	Number of layers
1	LSTM LSTM	LSTM	2
2	LSTM LSTM LSTM	LSTM	3
3	LSTM LSTM LSTM LSTM	LSTM	4
4	GRU GRU	GRU	2
5	GRU GRU GRU	GRU	3
6	GRU GRU GRU GRU	GRU	4

B. Inserting NA Values

NA Values were inserted in the three testing time series according to what is shown in Table II.

C. Implementation of Models

Once the time series and training data were selected, the first model was implemented, as shown in Fig. 5.

This model was trained with the data of the three-time series predicting 168 values for each time series.

Next, the remaining five models were implemented, predicting 168 values for each time series in every model. Fig. 6, Fig. 7, Fig. 8, Fig. 9 and Fig. 10 show the architecture of these models.

D. Evaluating Predictions

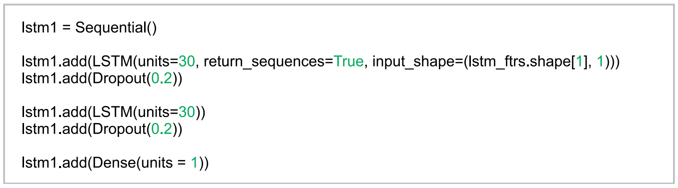
The results of six models are evaluated through Root Mean Squared Error (RMSE) according equation (4):

$$RMSE = \sqrt{\frac{\sum_{i=0}^{n-1}(Pi-Ri)^2}{n}} \quad (4)$$

The results achieved are described in the next section.

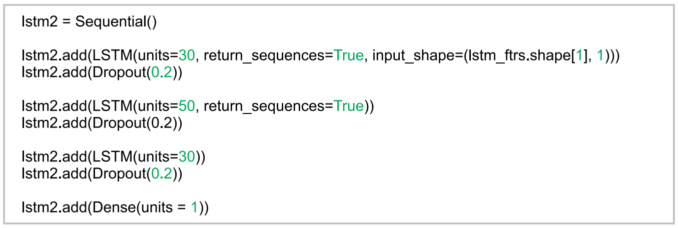
TABLE. II. Number of NA Values

	Temperature	Wind Direction	Wind Velocity
Short-Gaps	76	76	76
Medium-Gaps	124	124	124
Large-Gaps	155	155	155



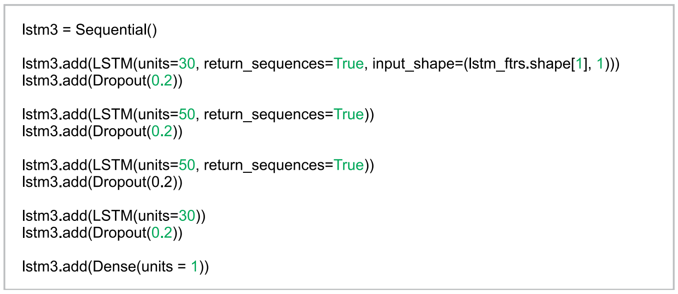
```
lstm1 = Sequential()

lstm1.add(LSTM(units=30, return_sequences=True, input_shape=(lstm_ftrs.shape[1], 1)))
lstm1.add(Dropout(0.2))

lstm1.add(LSTM(units=30))
lstm1.add(Dropout(0.2))

lstm1.add(Dense(units = 1))
```

Fig. 5. Architecture for First LSTM Model in Python.



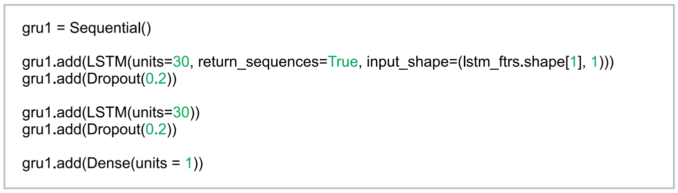
```
lstm2 = Sequential()

lstm2.add(LSTM(units=30, return_sequences=True, input_shape=(lstm_ftrs.shape[1], 1)))
lstm2.add(Dropout(0.2))

lstm2.add(LSTM(units=50, return_sequences=True))
lstm2.add(Dropout(0.2))

lstm2.add(LSTM(units=30))
lstm2.add(Dropout(0.2))

lstm2.add(Dense(units = 1))
```

Fig. 6. Architecture for Second LSTM Model in Python.



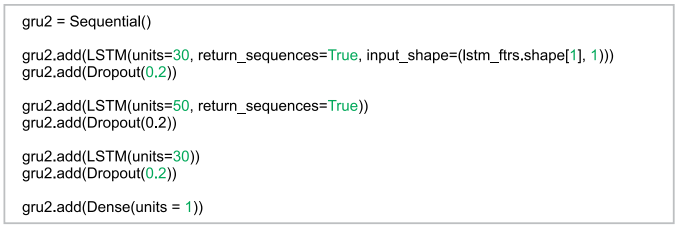
```
lstm3 = Sequential()

lstm3.add(LSTM(units=30, return_sequences=True, input_shape=(lstm_ftrs.shape[1], 1)))
lstm3.add(Dropout(0.2))

lstm3.add(LSTM(units=50, return_sequences=True))
lstm3.add(Dropout(0.2))

lstm3.add(LSTM(units=50, return_sequences=True))
lstm3.add(Dropout(0.2))

lstm3.add(LSTM(units=30))
lstm3.add(Dropout(0.2))

lstm3.add(Dense(units = 1))
```

Fig. 7. Architecture for Third LSTM Model in Python.



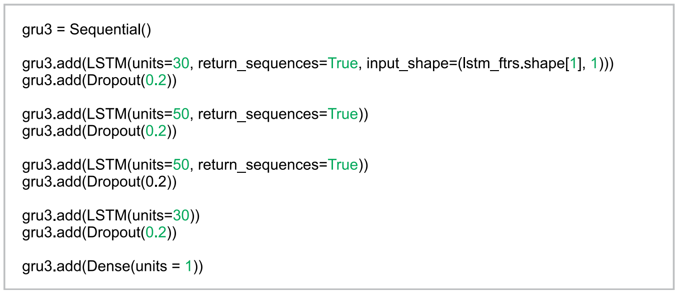
```
gru1 = Sequential()

gru1.add(LSTM(units=30, return_sequences=True, input_shape=(lstm_ftrs.shape[1], 1)))
gru1.add(Dropout(0.2))

gru1.add(LSTM(units=30))
gru1.add(Dropout(0.2))

gru1.add(Dense(units = 1))
```

Fig. 8. Architecture for First GRU Model in Python.



```
gru2 = Sequential()

gru2.add(LSTM(units=30, return_sequences=True, input_shape=(lstm_ftrs.shape[1], 1)))
gru2.add(Dropout(0.2))

gru2.add(LSTM(units=50, return_sequences=True))
gru2.add(Dropout(0.2))

gru2.add(LSTM(units=30))
gru2.add(Dropout(0.2))

gru2.add(Dense(units = 1))
```

Fig. 9. Architecture for Second GRU Model in Python.



```
gru3 = Sequential()

gru3.add(LSTM(units=30, return_sequences=True, input_shape=(lstm_ftrs.shape[1], 1)))
gru3.add(Dropout(0.2))

gru3.add(LSTM(units=50, return_sequences=True))
gru3.add(Dropout(0.2))

gru3.add(LSTM(units=50, return_sequences=True))
gru3.add(Dropout(0.2))

gru3.add(LSTM(units=30))
gru3.add(Dropout(0.2))

gru3.add(Dense(units = 1))
```

Fig. 10. Architecture for Third GRU Model in Python.

V. Results

This section shows the results achieved. Table III shows the corresponding RMSE values for imputation in hourly temperature time series.

According to what is shown in Table III and in Fig. 11 for the imputation process in the temperature time series on an average, the best model is LSTM LSTM LSTM (RMSE 0.5565), this model was also the one that produced the best results for all the gap-sizes.

Likewise, it can be seen that the GRU models were in second, third and fourth place, with the best GRU model being the GRU GRU (RMSE 0.5898) on average. So for this type of time series, the most recommended models would be the LSTM LSTM LSTM and the GRU GRU.

According to what is shown in Table IV and in Fig. 12 for the imputation process of wind direction time series, on average the best model was LSTM LSTM LSTM LSTM and this model was the best for each gap-size.

Likewise, it can be seen that similar to the temperature time series, the 4-layer LSTM model is the only one that managed to outperform the three GRU models and it is shown that the GRU models present more homogeneous results, while the LSTM models present results more heterogeneous that is, there is a greater dispersion among them.

According to Table V and Fig. 13 on average, the best model for imputation of the wind velocity time series was LSTM LSTM LSTM LSTM as well as for each gap-size.

TABLE. III. Temperature Time Series Results

Technique	RMSE			
	Short-Gaps	Medium-Gaps	Large-Gaps	Avg.
LSTM LSTM	0.7080	0.6934	0.6905	0.6973
LSTM LSTM LSTM	**0.5696**	**0.5592**	**0.5407**	**0.5565**
LSTM LSTM LSTM LSTM	0.5830	0.6191	0.6177	0.6066
GRU GRU	0.5848	0.5906	0.5942	0.5898
GRU GRU GRU	0.6095	0.5920	0.5781	0.5932
GRU GRU GRU GRU	0.5894	0.5943	0.5961	0.5933

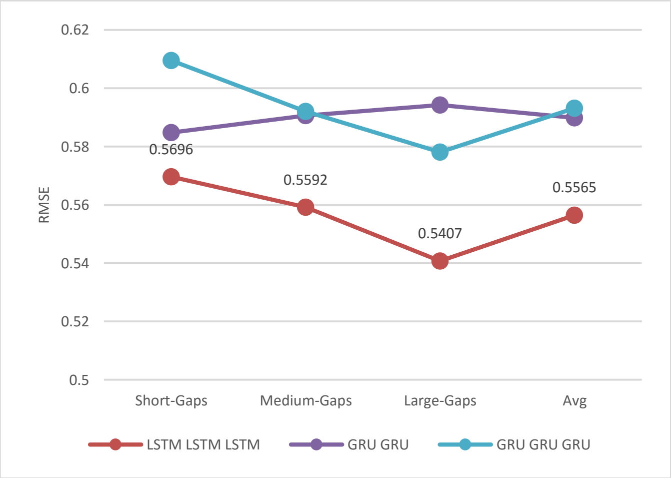


Fig. 11. Top 3 RNN Models for Temperature Time Series Imputation.

TABLE. IV. Wind Direction Time Series Results

Technique	RMSE			
	Short-Gaps	Medium-Gaps	Large-Gaps	Avg
LSTM LSTM	191.1086	190.4177	189.3457	190.2906
LSTM LSTM LSTM	186.5753	186.6744	185.5497	186.2664
LSTM LSTM LSTM LSTM	**150.4429**	**150.7271**	**151.1934**	**150.7878**
GRU GRU	161.7409	162.4947	161.2719	161.8358
GRU GRU GRU	163.6514	164.7290	164.7170	164.3658
GRU GRU GRU GRU	164.0155	164.2335	164.3080	164.1856

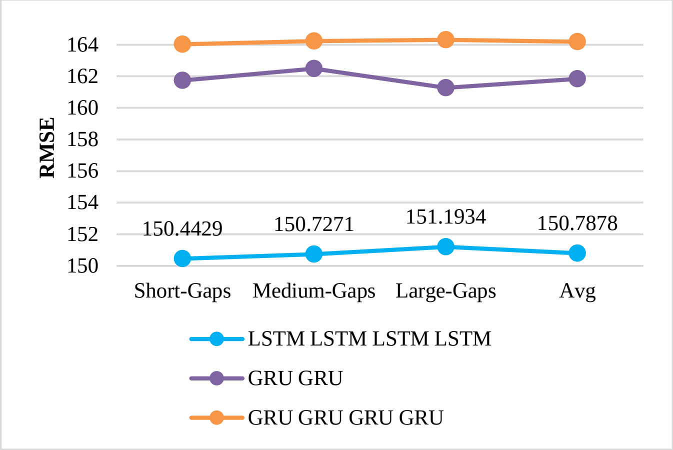


Fig. 12. Top 3 RNN Models for Wind Direction Time Series Imputation.

TABLE. V. WIND VELOCITY TIME SERIES RESULTS

Technique	RMSE			
	Short-Gaps	Medium-Gaps	Large-Gaps	Avg
LSTM LSTM	4.2175	4.2011	4.2451	4.2212
LSTM LSTM LSTM	2.7628	2.8512	2.8832	2.8324
LSTM LSTM LSTM LSTM	**2.3773**	**2.5066**	**2.5097**	**2.4645**
GRU GRU	3.5715	3.6008	3.6502	3.6075
GRU GRU GRU	3.6009	3.6276	3.6752	3.6345
GRU GRU GRU GRU	3.5501	3.5335	3.5599	3.5478

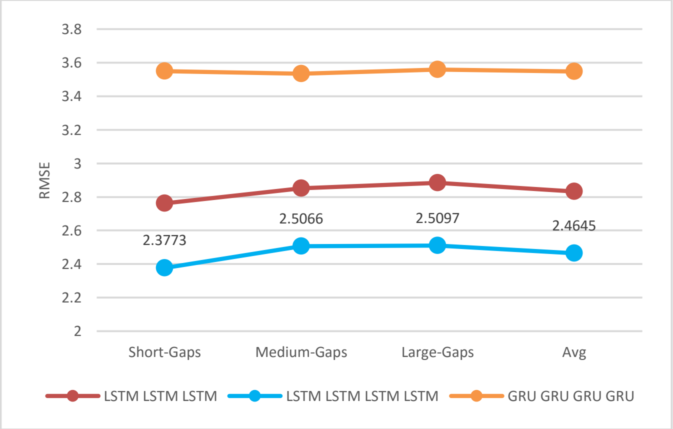


Fig. 13. Top 3 RNN Models for Wind Velocity Time Series Imputation

Unlike previous time series, for this type of time series two LSTM models present the best results: LSTM LSTM LSTM LSTM (RMSE 2.4645) and LSTM LSTM LSTM (RMSE 2.8324), while GRU models occupy the third, fourth and fifth place. Likewise, it is important to highlight that, like the previous time series, GRU models present more homogeneous results while LSTM models present more heterogeneous results.

VI. DISCUSSION

Next, the results achieved for the implemented models are compared with other univariate time series imputation techniques.

According to what is shown in Table VI, on average, the best technique for univariate time series imputation of temperatures is the recurrent neural network of 3 layers LSTM LSTM LSTM. However, performing an individual analysis for each gap-size, it is noted that this model is the best for medium-gaps (RMSE 0.5592) and large-gaps (RMSE 0.5407), but for short-gaps this is surpassed by the ARIMA-Kalman model (RMSE 0.4931).

According to Table VII, it is appreciated that on average the best technique for univariate time series imputation of wind directions is the recurrent neural network of 4 layers LSTM LSTM LSTM LSTM surpassing the different techniques based on moving averages in each gap-size.

According to Table VIII, it can be seen that on average the best imputation technique for wind velocity time series is the recurrent neural network of 4 layers LSTM LSTM LSTM LSTM. However, it is appreciated that this only exceeded the other techniques in large-gaps (RMSE 2.5097) while in Short-Gaps the best technique is Linear Weighted Moving Average (RMSE 1.1995) and for medium-gaps the best technique is Local Average of Nearest Neighbors (RMSE 1.6532).

TABLE. VI. TEMPERATURE IMPUTATION WITH ANOTHER TECHNIQUES

Technique	RMSE			
	Short-Gaps	Medium-Gaps	Large-Gaps	Avg
LSTM LSTM LSTM	0.5696	**0.5592**	**0.5407**	**0.5565**
GRU GRU	0.5848	0.5906	0.5942	0.5898
LANN	0.6343	2.0919	4.5210	2.4157
SMA	1.1208	1.8176	4.1515	2.3633
LWMA	0.8859	1.6868	3.9481	2.1736
EWMA	0.7262	1.6778	3.9348	2.1129
ARIMA-KALMAN	**0.4931**	1.13418	10.0067	3.8779

TABLE. VII. WIND DIRECTION IMPUTATION WITH ANOTHER TECHNIQUES

Technique	RMSE			
	Short-Gaps	Medium-Gaps	Large-Gaps	Avg
LSTM LSTM LSTM LSTM	**150.4429**	**150.7271**	**151.1934**	**150.7878**
GRU GRU	161.7409	162.4947	161.2719	161.8358
LANN	161.7666	169.3388	209.8176	180.3076
SMA	150.8792	153.2905	180.5009	161.5568
LWMA	150.4605	154.1056	184.2365	162.9342
EWMA	153.0640	161.0724	196.3272	170.1545
ARIMA-KALMAN	192.7941	233.8457	236.6198	221.0865

TABLE. VIII. WIND VELOCITY IMPUTATION WITH ANOTHER TECHNIQUES

Technique	RMSE			
	Short-Gaps	Medium-Gaps	Large-Gaps	Avg
LSTM LSTM LSTM LSTM	2.3773	2.5066	**2.5097**	**2.4645**
GRU GRU GRU GRU	3.5501	3.5335	3.5599	3.5478
LANN	1.2291	**1.6532**	4.0342	2.3055
SMA	1.4195	1.7944	3.6377	2.2838
LWMA	**1.1995**	1.6765	3.5758	2.1506
EWMA	1.2691	1.6852	3.5984	2.1842
ARIMA-KALMAN	1.2518	2.3963	3.3447	2.3309

As noted in the previous tables, the small difference between the RMSEs obtained by the models based on recurrent neural networks for short-gaps, medium-gaps and large-gaps should be highlighted. That is, the RMSE varies very little and it costs almost the same to impute 1 or 2 values than 30.

Likewise, it is also important to highlight that for short-gaps the imputation techniques of the state of the art offer very good results, while their performance is diminished in medium-gaps and much more in large-gaps, where RNN models offer the best results.

VII. CONCLUSIONS

The effectiveness of six models based on recurrent neural networks in nine case studies was analyzed, and in seven of them at least one model based on recurrent neural networks outperformed other imputation techniques of the state of the art, so we conclude that models based on recurrent neural networks are highly recommended to be implemented for univariate time series imputation especially for medium and large gap-sizes.

The results achieved show that not all models achieve optimal results, so it is important to implement not only one model but several in such a way that the most appropriate model can be chosen for the problem to solve.

In the three time series analyzed, the LSTM-based models show greater heterogeneity in their results compared to GRU-based models whose results are more homogeneous.

VIII. FUTURE WORK

In the present work it was experimented with models based on recurrent neural networks, differentiating them only by the number of layers and the number of neurons in each layer, for future works it would be important to be able to implement hybrid models that contain both LSTM and GRU layers, since it has been seen in different works that hybrid models for certain time series produce better results than non-hybrid models. Likewise, it can be experimented with other parameters such as the number of epochs, the batch-size, the training data size, the optimizer, etc.

Likewise, the results achieved by the RNN-based models for the wind direction and wind velocity time series, despite exceeding the state-of-the-art techniques, are not optimal (they have a high RMSE) so they could be improved by increasing the size of the training data or adding more variables to the model.

REFERENCES

[1] S. Moritz, T. Sardá, T. Bartz-Beielstein, M. Zaefferer and J. Stork, “Comparison of different methods for univariate time series imputation in R,” arxiv.org, 2015.

[2] A. Flores, H. Tito and C. Silva, “Local Average of Nearest Neighbors: Univariate Time Series Imputation,” International Journal of Advanced Computer Science and Applications, vol. 10, nº 8, pp. 45-50, 2019.

[3] S. Moritz, T. Bartz-Beielstein, “imputeTS: Time Series Missing Value Imputation in R,” The R Journal, vol. 9, nº 1, pp. 207-2018, 2017.

[4] R. Hyndman & G. Athanasopoulos, Forecasting: principles and practice, Melbourne, Australia: OTexts, 2018.

[5] Y. LeCun, Y. Bengio & G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436-444, 2015.

[6] C. Kyunghyun, V. Bart, G. Caglar, B. Dzmitry, B. Fethi, S. Holger & B. Yoshua, “Learning phrase representations using RNN enconder-decoder for statistical machine traslation,” arxiv.org, pp. 1-15, 2014.

[7] A. Flores, H. Tito and C. Silva, “Local average of nearest neighbors: univariate time series imputation,” International Journal of Advanced Computer Science and Applications, vol. 10, nº 8, pp. 45-50, 2019.

[8] A. Flores, H. Tito & C. Silva, “CBRm: case based reasoning approach for imputation of medium gaps,” (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 10, nº 9, pp. 376-382, 2019.

[9] A. Kowarick and M. Templ, “Imputation with R Package VIM,” Journal of Statistical Software, vol. 74, nº 7, 2016.

[10] M. Jahan Khan, H. Hayat and I. Awan, “Hybrid case-base maintenance approach for modeling large scale case-based reasoning systems,” Human-centric Computing and Information Sciences, vol. 9, nº 9, 2019.

[11] A. Flores, H. Tito & C. Silva, “CBRi: a case based reasoning-inspired approach for univariate time series imputation. In Press,” de 6th IEEE Latin American Conference on Computational Intelligence LA-CCI, Guayaquil, Ecuador, 2019.

[12] A. Flores, H. Tito & D. Centty, “Model for time series imputation based on average of historical vectors, fitting and smoothing,” (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 10, nº 10, pp. 346-352, 2019.

[13] R. Pascanu, T. Mikolov, Y. Bengio, “On the dificulty of training recurrent neural networks,” de 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013.

[14] M. Paco, C. López Del Alamo & R. Alfonte, “Forecasting of meteorological weather time series through a feature vector based on correlation,” de 18th International Conference Computer Analysis of Images and Patterns CAIP 2019, Salerno, Italy, 2019.

[15] W. Gail, G. Yoav & Y. Eran, “On the practical computational power of finite precision RNNs for language recognition,” arxiv.org, pp. 1-9, 2018.

Enhance the Security and Prevent Vampire Attack on Wireless Sensor Networks using Energy and Broadcasts Threshold Values

Hesham Abusaimeh
Associate Professor of Computer Science Department
Middle East University
Amman, 11831 Jordan

***Abstract*—Measuring and monitoring the surrounded environment are the main tasks of the most battery-based Wireless Sensor Networks (WSNs). The main energy consumption in the WSN is the communication and transferring data between these nodes. There are many researches works on how to preserve the energy consumption of the nodes inside this network. Most of these methods could save the energy and made the WSN lives for longer. However, there might be another reason of consuming energy and loosing these nodes from the network by the threats that targeting this kind of Networks such as the vampire attack that load the WSN with fake traffic. In this paper, a proposed method is presented of preventing the vampire attack from wasting the energy of the sensor nodes based on the energy level of the intermediate nodes in the way to the destination.**

Keywords—Network security; vampire attack; sensor nodes; energy; lifetime; power consumption; packets delivery ratio

I. Introduction

WSN is an ad-hoc nature, self-configured network in term of establishing its grid topology and distributed itself in the environment. This kind of network can be used in many applications such as environmental measurement, industrial, health and military applications. Therefore, protecting the network energy is very important such as the technique proposed in [1]. These sensor devices mainly gather the information from the physical environment, such as temperature degree and humidity, and send it by other nodes in the network until it reaches the destination node or the base station [2].

As WSNs, become more and more crucial to the everyday functioning of people and organizations, availability faults become less tolerable. In WSN, the energy is the most important factor since its energy coming from the attached non-chargeable battery. This energy is mainly consumed in gathering and forwarding information in the network. All the nodes in the flat wireless sensor nodes are communicating using peer-to-peer fashion without the need of central access point, which may cause many security threats to this kind of network [3] [4].

There are many security breaches that can affect the WSN, one of them is the vampire Attack, which will be described in Section II; Section III will explain some of the related work to prevent such kind of attacks, Section IV will give more details about the proposed method to prevent Vampire attack by modifying the PLGP protocol to consider the router energy and the broadcast average of each node. The paper will be concluded in Section V.

II. Vampire Attack

During the transmission of the data in the WSN after triggering some event in the environment, many attacks can affect this transmitted data and reduce the energy level of these sensors. These attacks aim to destroy the network by reducing the lifetime of each sensor node and prevent the delivery of the data packets [5].

One of these attacks that affect the data while it is transferred is the Vampire attack. There are two types of vampire attack, the Carousal attack and the Stretch attack. In the Carousal attack, the malicious node composes a fake packet to be transmitted by intermediate nodes in the WSN in certain path frequently. Repeating this fake packet by the malicious node in loop among the intermediate nodes will delay the service in the WSN. It also will increase the energy consumption and reduce the lifetime of the sensor nodes among this loop which will also reduce the lifetime among the WSN as shown in Fig. 1, where the fake packet is transmitted in loop among the nodes (A, B, C, and D) [5].

In the Stretch Attack as in Fig. 2, the path between the source node and the destination node will be stretched to include all the nodes in the network. This will be done during the forward process by a malicious node. Using this long path will consume more energy than the optimal path which is a lot shorter, therefore more nodes will have energy reduction and lose their lifetime [4]. Energy usage increase of factor O(min(N,ג)), where N is the number of nodes in the network and ג is the maximum path length allowed.

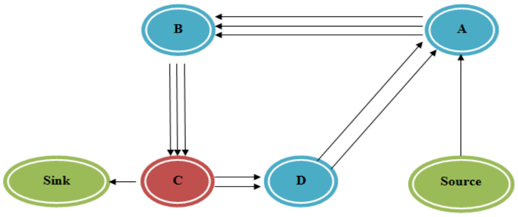


Fig 1. Carousel Attack [5].

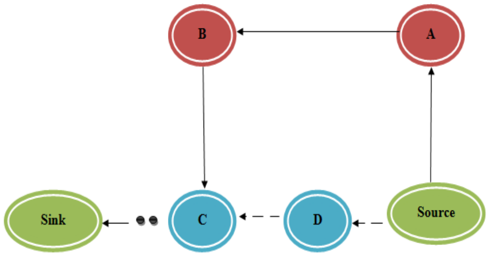


Fig 2. Stretch Attack [5].

Therefore, both famous types of vampire attack reducing the lifetime of the WSN either by having the packet to go in loop or to increase the number of intermediate nodes in the path to the destination.

III. Related Work

Nandhini, M. et al. proposed a method called Stop Transmit and Listen to detect the malicious node in the WSN. All nodes in the network will stop their transmission at a certain built-in time and listen for malicious transmission in the neighbor nodes, which can be malicious when it is sending in the non-transmitting time [6].

Pinky. B. et al. suggested content-based multicasting approach to defend the vampire attack and increase the battery life by enforcing a certain path of the data from the source to the destination [5].

Patil, A. and Giakwad, R. proposed a trust model system to prevent the vampire attack by defining a trust parameter for each sensor in the network [7].

Eugene Y. Vasserman and Nicholas Hopper introduced a definition for vampire attacks. Authours evaluates the vulnerabilities of existing protocols to routing layer energy depletion attacks. It is found that existing secure routing protocols such as Ariadne, SAODV, and SEAD does not provide security against Vampire attacks. The authors proposed defenses against some of the forwarding-phase attacks and introduced PLGP. The fully satisfactory solution for Vampire attacks during the topology discovery phase is not offered and further modifications to PLGPa is suggested [8].

Sivakumar and Murugapriya described the detection and elimination of vampire attacks in sensor networks. Authors proposed Optimal Energy Boostup protocol for providing the security. The PLGP protocol is performed as a tree structure. It is predicted that vampire attacks based on the behaviors of nodes and used to find the optimal path. It is found that the network energy is increased based on the location in forwarding phase [9].

In 2014, Menasinakai et al., explained the prevention and detection of vampire attacks. PLGP protocol is used to prevent the vampire attacks. To securely transmit the data, the path tracking technique is used in PLGP. The buffer technique also used in the proposed system in which the details of previous activity of every node is stored in a small buffer. It is found that, the proposed scheme is performed well to prevent attacks and achieved high energy consumption [10].

All of the above methods deal with the vampire attack as it is a serious problem in the WSN. Therefore, there is a need to detect this attack at early stages. In addition, most of these techniques considered their proposed method to be run over the clean-slate sensor network routing protocol (PLGP), which is originally considered as vulnerable to the vampire attack. PLGP has a discovery phase and then forwarding phase. There is also an enhancement on PLGP with attestations (PLGPa) by adding a verifiable path history to every PLGP packet. The addition of extra packet for verification also increases processor utilization, enquiring time and power consumption [11].

IV. Proposed Method and Simulation Results

Having vampire attack in any WSNs is critical. There is a need to detect these attacks as early as possible and prevent them from badly consuming the WSN energy level. The high availability of these networks is the critical property, and should stay alive even under malicious conditions. The PLGP protocol bounds damage from vampire attack, but it has some drawbacks. The PLGP includes path attestations, increasing the size of every packet results in increased bandwidth use, and thus radio power used to transmit these packets. In addition, of extra packet for verification, it also increases processor utilization, requiring time, and additional power for cryptographic computations and operations. PLGP is not considered to be vulnerable to Vampire attacks during the forwarding phase, but it might be vulnerable to that during the route discovery phase. Therefore, there is a need of new protocol that consumes less energy and discover the route and keep sending the data packets without any penetration by the vampire attacks. The modified PLGP protocol proposed in this paper to be enhanced version of PLGP and detect then prevent the vampire attacks in WSNs. The modified PLGP is based on reducing the processor utilization; requiring less detection time and maximizing the network lifetime. In this technique, we are arranging the sensor nodes in a particular manner so that there will be very less chances for occurrences of Vampire attack in the WSNs. This proposed modification of PLGP concentrate on measuring the route energy during the route discovery process, and it is also based on calculating the energy of the intermediate node on the route during the transmission of the data packet to the destination node. This route energy parameter will be calculated based on [1] route energy model and it will be added to the PLGP factors of choosing the best route to start routing the data packet.

Basically, the modified PLGP protocol detects the vampire attack based on two elements, the first one is the number of hops among the intermediate nodes, and the second one is the energy consumption of the routes to the destination. If any of these have been increased more than a threshold values expected of each route the discovery process will be initiated again by the modified PLGP. This regular checking and route discovery initiation will eliminate any suspicious nodes to be in the intermediate nodes in the route and choosing the shortest route without any loop in the middle.

In addition, the threshold value of the number of hops have been calculated based on the proposed model in [12]. This model discovers the malicious nodes during the discovery phase based on the average number of broadcast packets in the network as in the following equation:

$$Threshold = \sum_{i=1}^{n} \frac{Number\ of\ Broadcast}{N}$$

Moreover, the malicious node and the attacked route by the vampire attack will be detected based on the threshold value and the route energy of the route. Therefore, if the energy of the route is on high consumption speed and the node broadcast packet average is larger than the threshold value, then the network might be under vampire attack. Afterword, the source node will initiate the route discovery process of the PLGP again in order to choose new route and eliminate any malicious node participating in the route to the destination. This will make the current route to be the shortest route and eliminate any loop to the destination sensor node.

The modified PLGP also calculated the packet delivery ratio after transmitting the whole data in order to guarantee the enhancement on delivery percentage of each source node packets to the destination node.

The proposed modified PLGP routing protocol in the WSN has been tested by implementing the updated model of the route energy and the broadcast value in the network layer of the sensor devices Using NS-3 simulation environment. The simulation was conducted of grid of 100 sensor nodes and Personal Area Network (PAN) coordinator to establish the network and establish the traffic in the network. These nodes are distributed systemically as shown in Fig. 3. The simulation has been run of 60 minutes period with initial full battery of sensor energy of 40 Joules.

Fig. 4 shows the simulation scenario that is created to generate the data packet from the coordinator node to different destinations, and many vampire attacks were established from different malicious node in the WSN grid. Afterward the results have been captured of the modified PLGP routing protocol and compared that with the original PLGP in term of energy consumption and node lifetime. Finally, the delivery ratio of the sent packet has been calculated to in both protocols. The following figure also shows the data packet path from the coordinator node to various destination nodes in dashed lines.

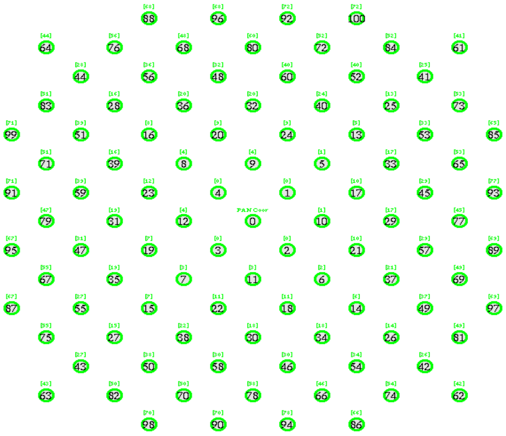


Fig 3. Sensors distribution in the Wireless Sensor Networks.

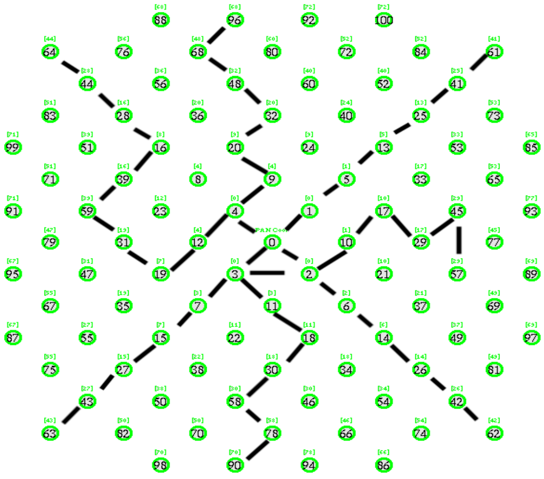


Fig 4. Traffic paths in the WSN grid.

Firstly, the consumption rate of the sensor node energy was compared between the modified PLGP with the route energy and number of broadcast approach to the tradition PLGP, where the vampire attack is most like to be happened and consume all the WSN energy. The results showed that the modified PLGP has reduced the energy consumption of each node in the WSN. Consequently, the energy consumption speed in the tradition PLGP was 0.03 Joules/Seconds, the modified PLGP has slowed the energy consumption average to 0.016 Joules/ Seconds, as in Fig. 5 shows the energy consumption rate of each wireless sensor nodes in both protocols.

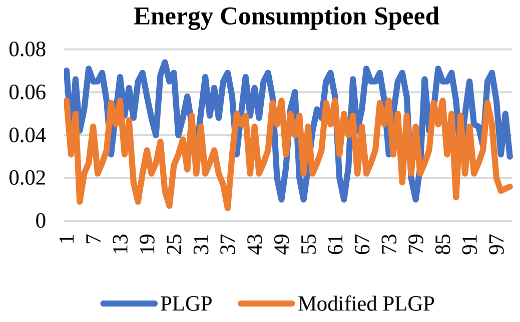


Fig 5. Energy Consumption Rate of the Sensor Nodes.

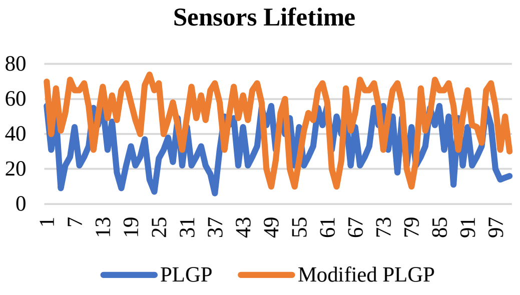


Fig 6. Wireless Sensor Nodes Lifetimes.

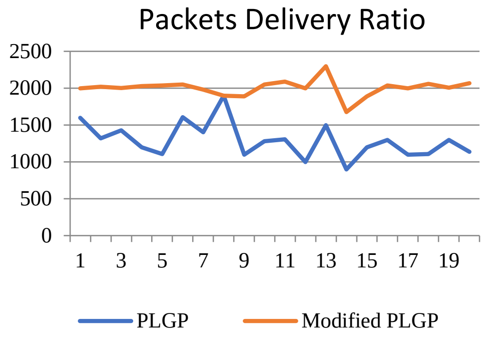


Fig 7. Packet Delivery Ratio of the Generated Packets.

The previous Fig. 6 clearly presents the results of the lifetime of each sensor nodes in the WSN among the Modified PLGP and the traditional PLGP. The results showed that the sensor node in the WSN that used the modified PLGP preserve more energy in the sensor nodes during transmission by detecting and preventing attack in the WSN. The average lifetime of the sensor node in the WSN that used the Modified was 54 minutes, where the lifetime of the sensor node in the WSN that used the traditional PLGP was 30 minutes.

The presence of the vampire attacks in the WSN decreases the delivery ratio to the destination node. The simulation results showed that the packet delivery ratio average of the generated packet was 2000 Bits Per Seconds (BPS), when the WSN used the modified PLGP. While, the packet delivery ratio when the WSN used the traditional PLGP was 1300 BPS as shown in Fig. 7.

V. Conclusion

WSN is used in many critical applications that are targeted from various malicious attacks. One of these is the vampire attack which has two kinds the carousal attack and stretch attacks. These vampire attacks consume all the power of the sensor nodes in the WSN and reduce the packet delivery ratio. The proposed technique has modified the PLGP protocol to consider the route energy of the intermediate nodes and the number of the broadcast packets of each node. In addition, the modified PLGP has initiated the discovery process of the PLGP and prevent and malicious node in the network from sending packets and prevent the loops in any route to the destination nodes by comparing the route energy and the number of the broadcast which should be less than the threshold value. The new proposed technique has been implemented using NS-3 simulation to compare the result of the modified PLGP with the original PLGP. All the simulation results showed that the modified PLGP has better performance in term of the nodes consumption speed and increasing the nodes lifetime. In addition, the packets delivery ratio is also calculated and compared between the two protocols, and the modified PLGP has also presented increasing the delivery ratio by 50% at the destination nodes.

Acknowledgment

The author is grateful to the Middle East University, Amman, Jordan for the financial support granted to cover the publication fee of this research article.

References

[1] H. Abusaimeh, M. Shkoukani and F. Alshrouf, "Balancing the Network Clusters for the Lifetime Enhancement in Dense Wireless Sensor Networks," Arabian Journal for Science and Engineering, 2013.

[2] A. Mahafzah and H. Abusaimeh, "Optimizing Power-Based Indoor Tracking System for Wireless Sensor Networks using ZigBee," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 9, no. 12, pp. 255-230, 2018.

[3] H. Abusaimeh, "Low Energy Consumption Rateinhome Sensors Using Prime Nodes," ARPN Journal of Engineering and Applied Sciences, vol. 13, no. 22, pp. 8738-8744, NOVEMBER 2018.

[4] V. Lokhande, S. DESHMUKH and S. SUTAR, "Vampire Attacks Prevention In Wireless Sensor Network," International Journal Of Current Engineering And Scientific Research (IJCESR), vol. 3, no. 1, 2016.

[5] P. Beaula, C. Anand and R. Gnanamurthy, "Defending Against Energy Draining Attacks in Wireless Sensor Networks with Secure Synchronization," International Journal of Science and Engineering Research (IJ0SER), vol. 3, no. 3, 2015.

[6] T. Sathyamorthi, D. Vijayachakaravarthy, R. Divya and M. Nandhini, "A Simple and Effective Scheme to find Malicious node in Wireless Sensor Network," International Journal of Research in Engg. And Tech., vol. 3, no. 2, 2014.

[7] A. Patil and R. Gaikwad, "Preventing Attack in Wireless Sensor Network by Using Trust Model," International Journal of Engineering Research & Technology (IJERT), vol. 4, no. 6, pp. 254-258, 2015.

[8] E. Y. Vasserma and N. Hopper, "Vampire Attacks: Draining Life from Wireless Ad Hoc Sensor Networks," IEEE Transactions on Mobile Computing, vol. 12, no. 2, Feb 2013.

[9] K. Sivakumar and P. Murugapriya, "Efficient Detection and Elimination of Vampire Attacks in Wireless Ad-Hoc Sensor Networks," International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, no. 1, 2014.

[10] S. D. SoumyashreeMenasinakai, "Prevention and Detection of Vampire Attacks Problem in Wireless Ad-Hoc Sensor Network," in International Conference on Information and Communications Security Protocols, Hong Kong, 2014.

[11] B. Parno, M. Luk, E. Gaustad and A. Perrig, "Secure sensor network routing: A clean-slate approach," in he 2006 ACM Conference on Emerging Network Experiment and Technology, Lisboa, 2006.

[12] D. Verma, G. Singh and K. Patidar, "Detection of Vampire Attack in Wireless Sensor Networks," International Journal of Computer Science and Information Technologies, vol. 4, no. 6, 2015.

Author's Profile

Hesham Abusaimeh is an associate professor of computer science in the Middle East University, Jordan. He is also senior IEEE member in the Jordan Section. Dr. Abusaimeh received his B.Sc. and M.Sc. Degrees both in computer science from Applied Science University and New York Institute of Technology respectively. Dr. Abusaimeh has received his Ph.D. from Loughborough University in the UK in 2009 in computer networks. His research interest includes wireless sensor networks, routing protocols, cyber security, and energy-aware routing protocols in sensor networks. Nowadays, Dr. Abusaimeh is the Dean of Graduate Studies and Scientific Research and the Dean of International Programmes at the Middle East University.

Automated Measurement of Hepatic Fat in T1-Mapping and DIXON MRI as a Powerful Biomarker of Metabolic Profile and Detection of Hepatic Steatosis

Khouloud AFFI[1], Mnaouer KACHOUT[2]
College of Science, Gafsa University, Gafsa, Tunisia[1]
Department of Computer Engineering, College of Computer Science and Engineering, University of Hail, KSA[2]
Innov'COM, Sup'Com, Carthage University, Tunis, Tunisia[2]

***Abstract*—Abnormal or excessive excess of intraperitoneal fat at different anatomical sites (heart, kidneys, liver, etc.) alters the metabolic profile by generating diseases causing cardiovascular complications. These include hepatic steatosis, which requires being increased surveillance before its severe progression to cirrhosis and its complications. Our objective in this study (invivo) was to propose a new approach to characterize and quantify hepatic fat. Then, differentiated patients with metabolic diseases, obesity, Type 2 diabetes (T2D), metabolic syndrome and healthy subjects. This distinction was not only according to traditional measurement tools such as body mass index (BMI) and waist circumference, but also according to the amount of fat from magnetic resonance imaging (MRI) DIXON image and T1-mapping at 1.5 Tesla (T). The evaluation results show that our proposed approach is reproducible, fast and robust. The distribution of the amount of hepatic fat in a cohort of data composed of four groups shows that hepatic fat is able to differentiate the metabolic population on the study chest. Relationship study of hepatic fat and cardiovascular parameters shows that hepatic fat is able to differentiate the metabolic population on the study chest. The relationship study of hepatic fat and cardiovascular parameters shows that hepatic fat has a negative influence on the heart if the amount it increases.**

Keywords*—Nonalcoholic fatty liver disease; non-alcoholic steatohepatitis; image processing; metabolic diseases; magnetic resonance imaging; active contour

I. Introduction

Nonalcoholic fatty liver disease (NAFLD) is one of the most common cause of chronic liver disease, its prevalence is rising worldwide and is estimated to affect 30% of adults and 10% of children in the United States [1]-[2]. Its rates are rising internationally alongside the growing epidemics of diabetes, obesity, and metabolic syndrome [3]–[4].

NAFLD is encompasses a range of liver histology severity in the absence of chronic alcohol use [5]-[6], it is commonly classified into two phenotypes, non-alcoholic fatty liver (NAFL) and non-alcoholic steatohepatitis (NASH). The most form is simple steatosis in which triglyceride accumulates within hepatocytes. A more advanced form of NAFLD, nonalcoholic steatohepatitis, includes inflammation and liver cell injury. The development of NASH is associated with an increased risk for morbidity and mortality through hepatic (fibrosis, cirrhosis, hepatocellular carcinoma) and non-hepatic (cardiovascular disease and cancer) complications [7]–[8].

The underlying cause of NAFLD, insulin resistance, leads to intracellular accumulation of triglycerides in hepatocytes (steatosis) [9]-[10]. Currently, therapeutic trials in NASH require medical imaging techniques that have greatly contributed to the detection of liver steatosis such as biopsy, ultrasound, computed tomography (CT) and recently Magnetic Resonance Imaging (MRI). The gold standard for establishing diagnosis as well as severity of NAFLD is liver biopsy, but it is invasive, poor patient acceptance, requires of a hospitalization, not exempt of complications and suffers from tremendous sampling variability [11].

Diana Feier, Ahmed Ba-Ssalmah and al. estimated that only a tiny fraction of the liver (roughly *1/50.000*), leading to sampling errors [12]. They also showed that liver biopsy samples contain at least 11 portable triads and measure at least *2.0* cm to reduce sampling variability. Other studies have also shown similar sampling variability [13]-[14].

Therefore, in clinical practice, ultrasound is often used to assess NAFLD. However, the lack of sharpness due to noise limited its role in the classification of the degree of steatosis [15]-[16].

Parficio and al. [5] classified ultrasound images were graded independently for presence and the severity of steatosis by two radiologists. Steatosis was defined by an appearance of hepatic parenchymal in which the liver was considered be normal if there was normal liver echo texture with clear visualization of the internal vascular system. The severity of steatosis was classified as mild, moderate or severe according to previously defined criteria. In fact, mild steatosis was recognized by a slight increase in the echogenicity of the liver parenchymal or no posterior beam attenuation.

Severe steatosis was recognized by coarsely increased hepatic parenchymal echotexture and subsequently marked beam attenuation. Moderate steatosis was recognized by ultrasound characteristics of liver ultrasound texture, and beam attenuation between light and severe parameters.

CT is an x-ray imaging technique, Given its imprecision in detecting mild hepatic steatosis and potential radiation risk, computed tomography is not suitable for the evaluation of hepatic steatosis in the general population, but can be effective in specific clinical situations, such as assessing donor candidates for liver transplantation [17].

Disease assessment within clinical practice for NAFLD is currently done with MRI. In contrast to other imaging techniques such as ultrasound and computed tomography, which use proxies to assess hepatic steatosis (i.e., attenuation and echogenicity), Seung Soo Lee and Seong Ho Park in [17] and Parambir S.Dulai, Claude B.Sirlin, Rohit Loomba in [18] shows that magnetic resonance spectroscopy (MRS) and magnetic resonance imaging are the most accurate and reliable methods of quantifying liver fat. In our article, we are focused on magnetic resonance imaging.

Several MRI methods have been introduced to quantify hepatic fat, including chemical-shift imaging (CSI) to differentiate protons in fat from those in water, that is, the difference in MRI frequency between protons in fat and water [19]. Other methods used fat saturation, and fat-selective excitation approaches [20]-[21]. The CSI approach is most widely used because of its easy applicability and higher accuracy. Indeed, CSI techniques separate magnetic resonance (MR) signals into water (W) and fat (F) components based on the chemical shift between fat and water.

The diagnosis of fatty liver often involves the use of conventional measurement tools, But these methods remain inappropriate, therefore, other criteria must be taken into diagnostic hepatic steatosis, MRI, non- invasive examination, provide multi parametric information, a high-resolution image with an absence of completely harmless radiation.

In this paper we will propose a new approach for non-invasive quantification of intraperitoneal fat. Therefore, we wish to evaluate our study on cohort of data composed of four groups in order to prove that hepatic fat is able to differentiate patients with metabolic diseases' obesity, T2D, metabolic syndrome and healthy subjects. This distinction is according to the amount of fat from segmentation of MRI DIXON and T1-mapping images at 1.5T in the first heading. In the second heading, this study investigates the relationship between the correlation of hepatic fat and cardiovascular disorders. Finally, we will predict cardiovascular complications for these patients.

II. Proposed Solution

To quantify the hepatic fat, we treated the water cards obtained from a specific DIXON sequence; Indeed Dixon imagery is based on the chemical displacement between water and fat protons, in order to separate their signal. In-Phase (IP): the total signal corresponds to the water signal to which is added that of the fat like Eq. (1). Out-of-Phase (OP): in Eq. (2) the total signal corresponds to the water signal from which the fat signal subtracts. It consists of making two spin echo acquisitions: the first for which the water and fat protons are in phase and the second signal for which the water and fat protons are out of phase. In Eq. (3) by adding the two signals, only that of water is displayed; by subtraction, we erase the water signal in favor of that of fat as presented in Eq. (4), results are shown in Fig. 1.

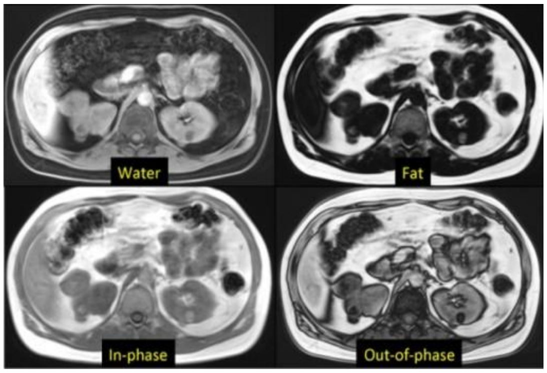


Fig. 1. Dixon Imaging.

Dixon imaging is based on the chemical displacement between water and fat protons, in order to separate their signal.

$$IP = W + F \tag{1}$$

$$IP = W – F \tag{2}$$

$$½[IP + OP] = ½\,[(W + F) + (W - F)] = ½\,[2W] = W \tag{3}$$

➡ Water only image

$$½[IP - OP] = ½\,[(W + F) - (W - F)] = ½\,[2F] = F \tag{4}$$

➡ **Fat only image**

III. Solution

MRI acquisitions including DIXON imaging were performed at 1.5T in 117 individuals (60 women, 50 men, age 47.5 ds): 15 obese patients, 25 metabolic syndrome patients, 40 type 2 diabetes patients and 19 healthy controls. 40 axial slices with 3 mm thickness, and in-plane resolution of 1.18 mm were acquired for each subject using a two-point Dixon sequence.

The segmentation process developed for liver fat quantification consists of three steps.

A first preprocessing step is needed to improve the quality of the MRI-DIXON-water map image. The second step is to segment the liver by combining different image processing methods (active contour Federal Trade Commission (FTC) with a double cycle of smoothing and regulation, a K-means machine learning method and mathematical morphology). The last step is to classify the liver into three classes. A class corresponds to liver fat, a class contains the vessels are presented in Fig. 2.

A. Pretreatment

To get better liver segmentation, the pretreatment step is essential before the segmentation process. In this step, we proposed the use of a morphological filter called “top hat”. The principle consists in calculating the opening of the image by a very specific structuring element then to subtract the result obtained from the original image. The morphological operation ‘opening’ consists in eroding the image followed by dilation by the same structuring element. Using the top hat in this study makes it possible to fill in the holes and correct the intensity inconsistency in the T1 card.

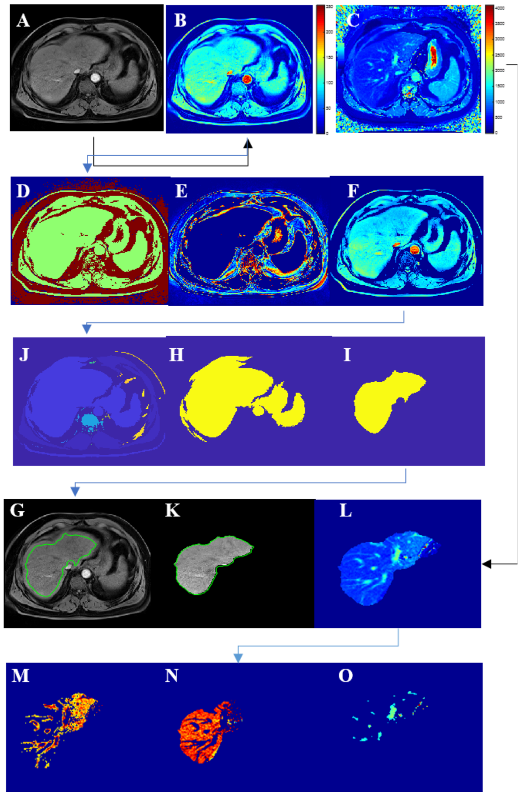


Fig. 2. Diagram of the Proposed Approach to Liver Segmentation: A: Fat Map; B: Pretreatment; C: T1-Mapping;D: Classification of the Water Map Into 2 Classes; E: Background of the Image; Liver Class; J: Distribution of the Image into Objects; H: the Largest Object Area; Mathematical Morphology Applied on (H); G: Active Contour on the Water Map; L: Superposition of the Liver Obtained on the T1-Mapping; M: Vessel; N: Liver Fat; O: Partial Volume.

Thus, we normalize and enhance the contrast of the image by adjusting the initial histogram values. This increases the contrast of the fat in relation to the background of the image and eliminates the shadow effects linked to the acquisition artifacts and to always have the same threshold.

B. Normalization

Normalization of an image consists in dividing each value of the histogram by the total number of pixels of the image to obtain a normalized histogram. This histogram corresponds to an empirical probability distribution (all values are between *0* and *255*). In Eq. (5) the formula used is as follows:

$$I(x,y)=255*I(x,y)/max(I(x,y)) \quad (5)$$

With *I* the original image, *x* and *y* the coordinates of a pixel.

C. Adjustment

Is to increase the contrast of the image, we were able to better distinguish fatty liver (in cases). The MATLAB function used is “Imadjust” with the default parameters. Thanks to this function, the image contrast is enhanced and the liver is more visible.

D. Liver Segmentation

This step consists in segmenting the liver from a water card obtained by the DIXON method by K-means (three classes) and choosing the liver class. To do this we distribute the image into objects. We get a card in which each object is labeled. The MATLAB function used is “bwlabel” with the default parameters (4 connections). Then the object with the largest area (the first part of the fat) as shown in Fig. 2 will be removed from the image. It is this object which corresponds to the mark of the liver.

Then, we used the mask of this object as an initialization of FTC and finally superimpose the mask obtained on the T1-mapping card to have only the liver and quantify the hepatic fat. The principle of an active contour consists in positioning the image, more precisely in the vicinity of the shape to be detected, an initial contour which will undergo a deformation under the effect of several forces such as: An internal energy *E* internal allowing to regularize the contour a potential energy *E* image linked to the image; An external energy *E* external linked to the particular constraints that can be added. These energies will allow the active contour to evolve to explore the minimum energy position which will thus be an arrangement between the various constraints of the problem.

E. Quantification of Hepatic Fat

For the fat quantization part, we classify the pixels of the time by the k-means algorithm into two classes.

K-means is a data partitioning algorithm (the pixels of the image in our case). The principle is as follows: given a set of pixels in the image $(x1, x2,\ldots, xn)$, we seek to partition the n pixels into sets $S = \{S1, S2,\ldots, Sk\}$ $(k \leqslant n)$ by minimizing the distance between the pixels inside each partition. In this case, the number of partitions is equal to two. The formula used to quantify hepatic fat from T1-mapping and DIXON- MRI images is the average of all the pixels of the fat class.

For the evaluation of my approach, metrics will be calculated to demonstrate the strengths of my application, such as inter / intra-operator reproducibility, the Dice coefficient, the coefficient of variation (CV) and the segmentation speed.

IV. Result and Discussion

Our objective in this study (in-vivo) is to propose a new approach to characterize and quantify hepatic fat and differentiate patients with metabolic diseases' obesity, T2D, metabolic syndrome and healthy subjects not only according to conventional measurement tools but also according to the amount of fat from the segmentation of DIXON MRI images and T1-mapping at 1.5 T. This first step should then make it possible to predict and assess the cardiovascular risks in these patients.

As a result, in Fig. 3, we evaluate the distribution of the quantity of hepatic fat on a cohort of data made up of four groups metabolic syndromes (n = 31), obese (n = 10), T2D (n = 48) and healthy subjects (n = 19).

P: value for quantifying the statistical significance of a result under a hypothesis

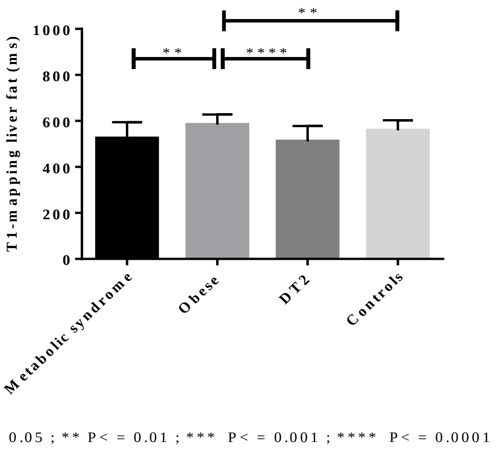


Fig. 3. Distribution of Hepatic Fat.

In the study of liver fat distribution in a population of four patient groups, significant differences were reported. The biggest difference was between obese and type 2 diabetics.

In Fig. 3, we also observe significant differences between metabolic and obese syndromes and between obese and control.

A. Assessment Metric

To assess the robustness and reproducibility of our segmentation approach, we calculated the intra-operator reproducibility by repeating the segmentation process on all groups of patients; it is defined as the absolute difference between two measurements divided by the average of two measurements. For inter-operator reproducibility, an operator who had not previously read patient data segmented the fat maps of the two groups of patients.

1) Inter-operator reproducibility: For patients with metabolic syndrome, the correlation between the fat measurements of the two operators is very good as presented in Table I. the Pearson correlation coefficient is 0.99. The inter-operator reproducibility is 4.9 ± 0.29%. For the obese, the correlation between the fat measurements of the two operators is also very good. We obtained a Pearson correlation coefficient of 0.98. The inter-operator reproducibility is 3.9 ± 0.20%. For type 2 diabetics, the correlation between the two operators' fat measurements is good. We obtained a Pearson correlation coefficient of 0.97. The inter-operator reproducibility is 8.3 ± 2.22%. Concerning healthy subjects, the correlation between the fat measurements of the two operators is good. We obtained a Pearson correlation coefficient of 0.963. The inter-operator reproducibility is 5.7 ± 0.16.

2) Intra-operator reproducibility: For metabolic syndromes, we obtained a Pearson correlation coefficient of 0.998 as shown in Table II. between the two measures. The intra-operator reproducibility is 3.5 ± 1.08%. Concerning T2D, we obtained a Pearson correlation coefficient of 0.973.

3) The intra-operator reproducibility is 1.1 ± 0.04%. For the obese, the correlation between the two measures is very good. The Pearson correlation coefficient is 0.999. The intra-operator reproducibility is 1.5 ± 0.11%. As for healthy subjects, we obtained an excellent correlation. The Pearson correlation coefficient is 0.999. The intra-operator reproducibility is 1.6 ± 0.075%.

4) Coefficient of variation: we calculated the coefficient of variation for each group, I always obtained values lower than 14% as presented in Table III.

5) Dice index: we also calculated the Dice coefficient as shown in this Table IV.

6) Speed of the segmentation: As for the speed of segmentation, our approach is very fast in the segmentation of the liver and the quantification of hepatic fat, it allows the detection of hepatic steatosis in a time less than 5 s / section as presented in Table V.

TABLE. I. INTER-OPERATOR REPRODUCIBILITY: CORRELATION BETWEEN THE FAT MEASUREMENTS OF THE TWO OPERATORS ON PATIENTS WITH METABOLIC SYNDROME AND OBESE PATIENTS STYLES

Groups	**Pearson Correlation Coefficient**
Metabolic syndromes	0.994
Obese	0.986
Type 2 diabetes	0.974
Control	0.963

TABLE. II. INTRA-OPERATOR REPRODUCIBILITY: CORRELATION BETWEEN THE FAT MEASUREMENTS OF THE TWO OPERATORS ON PATIENTS WITH METABOLIC SYNDROME AND OBESE PATIENTS

Groups	**Pearson Correlation Coefficient**
Metabolic syndromes	0.998
Obese	0.999
Type 2 diabetes	0.973
Control	0.999

TABLE. III. COEFFICIENT OF VARIATION FOR PATIENTS WITH METABOLIC SYNDROME, OBESE, T2 AND HEALTHY SUBJECTS

Groups	**CV (%)**
Metabolic syndromes	13.388
Obese	7.512
Type 2 diabetes	13.006
Control	7.905

TABLE. IV. DICE INDEX %

Groups	**Liver Fat%**
Metabolic syndromes	0.988
Obese	0.974
Type 2 diabetes	0.969
Control	0.985

TABLE. V. SPEED OF SEGMENTATION OF OUR APPROACH: THE METHOD IS VERY FAST

Segmentation speed (s/slice)				
Groups	Liver from Dixon	Liver from T1-mapping (s)	Liver Quantification (s)	Total time
Metabolic syndrome	2.45	1,87	0.780	5.10
Obese	2.53	1,66	0.877	5.06
Type 2 diabetes	2.35	1,87	0.890	5.11
Controls	2.7	1,98	1.020	5.70

B. *Statistical Study*

In this part, risk factors will be correlated with liver fat to study the metabolic links that may exist.

1) Relationship between liver fat and age: In the statistical study, we observed a significant relationship between the amount of fat and age. The amount of fat increases significantly with age for patients with metabolic syndrome (*P=0.04*) and obese patients (*P=0.045*) and also type 2 diabetics (*P=0.033*).

As for the holy subject this quantity has no significant connection with age (*P=0.54*) as shown in Fig. 4.

2) Relationship between liver fat and BMI: We also studied the correlation between the amount of fat and the body mass index. Indeed, the amount of fat tends to increase with BMI but not significantly for all groups, namely patients with metabolic syndrome (*P = 0.28*), obese (*P = 0.55*), type 2 diabetics (*P = 0.08*) and healthy subjects (*P = 0.89*) as shown in Fig. 5.

3) Relationship between liver fat and BSP: The amount of fat increases significantly with systolic brachial pressure for patients with metabolic syndrome (*P = 0.049*). However, it did not significantly for type 2 diabetics (*P = 0.9*) as shown in Fig. 6.

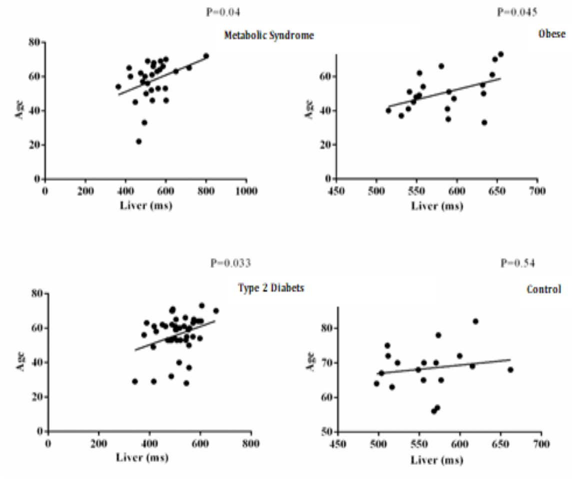


Fig. 4. Change in the Amount of Liver Fat as a Function of Age.

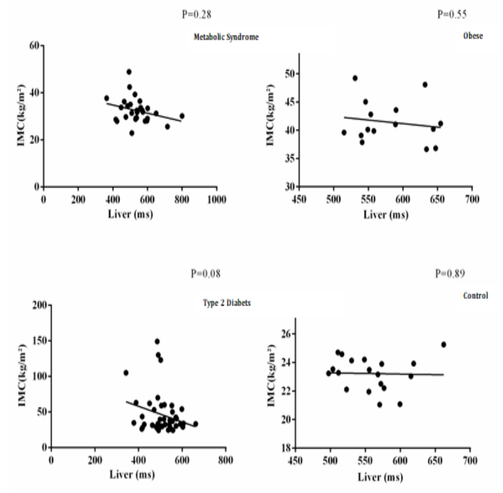


Fig. 5. Liver Fat and BMI.

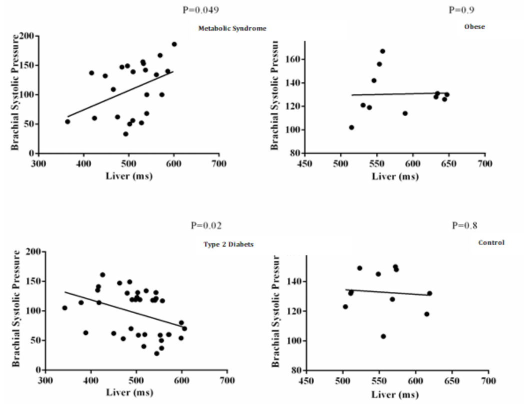


Fig. 6. Liver Fat and Brachial Systolic Pressure.

4) Relationship between liver fat and BDP: The amount of fat increases significantly with systolic brachial pressure for patients with metabolic syndrome (*P = 0.01*). However, it significantly reduced for type 2 diabetics (*P = 0.04*) as you can see in Fig. 7.

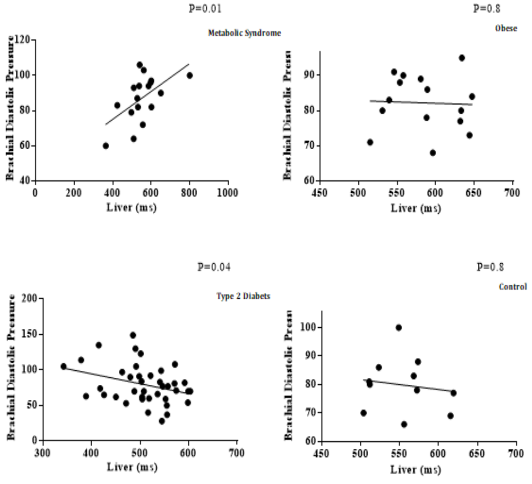


Fig. 7. Liver Fat and Brachial Dystolic Pressure.

V. Conclusion

In conclusion, the precise classification and quantification of hepatic fat is crucial for metabolic studies and the detection of fibrosis where they serve as good indicators of the associated metabolic and cardiovascular disorders. They can serve as an effective and precise tool for the diagnosis and differentiation of risk profiles of patients with metabolic diseases and could be considered in the future to predict cardiovascular complications. Relationship study of hepatic fat and cardiovascular parameters shows that hepatic fat has a negative influence on the heart if the amount it increases.

The perspectives of this work are many; first we want to segment the liver by deep learning algorithms. Then we also want to detect and quantify hepatic fibrosis from the T1 mapping.

References

[1] Tsiplakidou, M., Tsipouras, M. G., Manousou, P., Giannakeas, N., & Tzallas, “Automated Hepatic Steatosis Assessment through Liver Biopsy Image Processing”, IEEE 18th Conference on Business Informatics (CBI), 2016; DOI:10.1109/cbi.2016.51.

[2] Z.M. Younossi, Deirdre Blissett, Robert Blissett, Linda Henry, Maria Stepanova, Andrei Racila, “The economic and clinical burden of nonalcoholic fatty liver disease in the United States and Europe”, Hepatology, 2016; vol. 64, no. 5, pp. 1577-1586.

[3] Loomba R, Sanyal AJ, “The global NAFLD epidemic”, Nat Rev Gastroenterol Hepatol, 2013;10:686–690.

[4] Rinella ME, “Nonalcoholic fatty liver disease: a systematic review”, JAMA, 2015;313:2263–2273.

[5] Lucia Pacifico, Michele Di Martino, Carlo Catalano, Valeria Panebianco, Mario Bezzi, Caterina Anania, and Claudio Chiesa, “T1-weighted dual-echo MRI for fat quantification in pediatric nonalcoholic fatty liver disease”, World Journal of Gastroenterology: WJG, 2011;17: 30129.

[6] Spengler EK, Loomba R, “Recommendations for diagnosis, referral for liver biopsy, and treatment of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis”, Mayo Clin Proc, 2015;90:1233–1246.

[7] Chalasani N, Younossi Z, Lavine JE, Diehl AM, Brunt EM, Cusi K, Charlton M, Sanyal AJ, “The diagnosis and management of non-alcoholic fatty liver disease: practice guideline by the American Gastroenterological Association, American Association for the Study of Liver Diseases, and American College of Gastroenterology”, Gastroenterology,2012;142:1592–1609.

[8] Arulanandan A, Ang B, Bettencourt R, Hooker J, Behling C, Lin GY, Loomba R.,”Association between quantity of liver fat and cardiovascular risk in patients with nonalcoholic fatty liver disease independent of nonalcoholic steatohepatitis”, Clin Gastroenterol Hepatol, 2015;13:1513–1520.

[9] A. Han, J. W. Erdman, D. G. Simpson, M. P. Andre and W. D. O'Brien, “Early detection of fatty liver disease in mice via quantitative ultrasound”, IEEE International Ultrasonics Symposium, Chicago, IL, 2014; pp. 2363-2366.

[10] J. Kong, J. Lee Michael, P. Bagci, P. Sharma, D. Martin, N. Volkan Adsay, J. H. Saltz and A. B. Farris, “Computer-based Image Analysis of Liver Steatosis with Large-scale Microscopy Imagery and Correlation with Magnetic Resonance Imaging Lipid Analysis”, IEEE [International Conference on Bioinformatics and Biomedicine], 2011; 978-0-7695-4574-5/11.

[11] Bonekamp S, Tang A, Mashhood A, S Michael S. Middleton MD , “Spatial distribution of MRI Determined hepatic proton density fat fraction in adults with nonalcoholic fatty liver disease”, J Magn Reson Imaging, 2014;39:1525–1532.

[12] Diana Feier, Ahmed Ba-Ssalamah, “Current Noninvasive MR-Based Imaging Methods in Assessing NAFLD Patients”, IntechOpen, May 31st 2019; DOI: 10.5772/intechopen.82096.

[13] T. Nguyen, A. Podkowa, R. J. Miller, M. L. Oelze and M. Do, “In-vivo study of quantitative ultrasound parameters in fatty rabbit livers”, IEEE International Ultrasonics Symposium (IUS), Washington, DC, 2017; pp. 1-4.

[14] Harald Kramer, Perry J. Pickhardt, Mark A. Kliewer, Diego Hernando, Guang-Hong, “Accuracy of liver fat quantification with advanced CT MRI and Ultrasound techniques: prospective comparison with MR spectroscopy”, AJR Am. J Roentgenol, 2017; vol. 208, no. 1, pp. 92-100, Jan.

[15] Aiguo Han ; Andrew S. Boehringer ; Yingzhen N. Zhang ; Vivian Montes ; Michael P. Andre, “Improved Assessment of Hepatic Steatosis in Humans Using Multi-Parametric Quantitative Ultrasound”, IEEE International Ultrasonics Symposium (IUS), Glasgow, United Kingdom, 2019; pp. 1819-1822.

[16] Che-Chou Shen ; Sheng-Chang Yu ; Chia-Yuan Liu, “Using high-frequency ultrasound statistical scattering model to assess Nonalcoholic Fatty Liver Disease (NAFLD) in mice”, International Conference on Telecommunications and Signal Processing (TSP), Vienna, 2016; pp. 379-382.

[17] Seung Soo Lee and Seong Ho Park, “Radiologic evaluation of nonalcoholic fatty liver diseas”, World J Gastroenterol, 2014 Jun 21; 20(23): 7392–7402.

[18] Parambir S.Dulai, Claude B.Sirlin, Rohit Loomba, “MRI and MRE for non-invasive quantitative assessment of hepatic steatosis and fibrosis in NAFLD and NASH: Clinical trials to clinical practice”, Journal of hepatology, 2016; vol. 65: 1006–1016.

[19] Reeder SB, Cruite I, Hamilton G, Sirlin CB, “Quantitative Assessment of Liver Fat with Magnetic Resonance Imaging and Spectroscopy”, J Magn Reson Imaging, 2011 Oct; 34(4):729-749.

[20] J. Cui, B. Ang, W. Haufe, C. Hernandez, E. C. Verna, C. B. Sirlin, R. Loomba, “Comparative diagnostic accuracy of magnetic resonance elastography vs. eight clinical prediction rules for non-invasive diagnosis of advanced fibrosis in biopsy-proven non-alcoholic fatty liver disease: a prospective study”, Aliment Pharmacol Ther, 2015;41: 1271–1280.

[21] Rohit Loomba, Tanya Wolfson, Brandon Ang, Jonathan Hooker, Cynthia Behling, Michael Peterson, “Magnetic resonance elastography predicts advanced fibrosis in patients with nonalcoholic fatty liver disease: a prospective study”, Hepatology, 2014;60:1920–1928.

Producing Standard Rules for Smart Real Estate Property Buying Decisions based on Web Scraping Technology and Machine Learning Techniques

Haris Ahmed*[1], Tahseen Ahmed Jilani[2], Waleej Haider[3], Syed Noman Hasany[4]
Mohammad Asad Abbasi[5], Ahsan Masroor[6]
Department of Computer Science, Sir Syed University of Engineering and Technology, Karachi, Pakistan[1, 3, 4, 5, 6]
Department of Computer Science, University of Karachi, Karachi, Pakistan[2]
School of Computer Science, University of Nottingham, Nottingham, UK[2]

Abstract—**Purchasing of real estate property is a stressful and time-consuming activity, regardless of the individual in question is a buyer or seller. The act is also a major financial decision which can lead to numerous consequences if taken hastily. Therefore, it is encouraged that a person should properly invest their time and money in research relating to price demands, property type and location, etc. It can be a difficult task to assess what real estate property can be considered as the best property to buy. The key idea of the current research study is to create a set of standard rules, which should be embraced to make a smart decision of buying real estate property, based on web scraping technology and machine learning techniques.**

Keywords—Web scraping technology; HtmlAgilityPack; machine learning; C4.5 decision tree; Weka-J48

I. Introduction

Any decision in relation to a property purchase or sales is a vital decision. To say that it is difficult to make up one's mind in that circumstance is an understatement [1]. However, that is not to say as it is impossible to do so as there are technological means available to the modern man that allow them to make the best decision. One such route is to take the assistance of web scraping technology. This form of tech allows the user to find various online real state property advertisements from different web sources [2]. Therefore, the individual will have a much better idea of what sort of decision they should be making in terms of selling or buying real estate. Furthermore, with the help of machine learning techniques such as decision tree C4.5 [3], in combination with the prior mentioned option, one can easily make a superior decision.

II. Web Scraping using HTML Agility Pack

The term "Web Scraping" also referred to as the "screen scraping or web data extraction technique" is a program for mining huge volume of data from an internet source, removing the information and saving it to a local file in a computer or databank and it saves the table in a spreadsheet format [4].

The data exhibited on numerous internet sources can only be observed through a web browser. Therefore, the sole possibility is to physically copy-paste the information. This is a very monotonous task that can take a lot of time, even days to complete. In addition to this, web crawling is a procedure that mechanizes this process. As a result, Web Scraping software does not need to manually copy data from a source but can perform the same task quickly.

A. HTML Agility-Pack

This is a responsive "HTML parser" inscribed in C# that builds a read/write "DOM" and supports basic "XPATH" or "XSLT" [5]. It is a ".NET code library" that permits you to analyse "out of the web" HTML archives. For improved understanding, "HTML Agility pack" is used to contrivance scraping of several web pages present on the internet [6].

- HTML Parsing

HTML parsing is fundamental as taking in HTML code and mining applicable data like the title of the page, subsections in the page, relations, bold text etc.

- Document Object Model

The "Document Object Model" is a software design "API" for "HTML" and "XML" documents. It outlines the rational construction of documents and the method by which a document is retrieved and deployed [7].

B. HTML Agility Pack Installation Steps

1) First, install the "NuGet package".

2) Below the segment "Package Manager" copy the installed code. Such as, if there is a statement of "PM> Install-Package HtmlAgilityPack - Version 1.5.1," then the text following the "PM>" shall be copied by the user.

3) Afterwards, go to the "Visual Studio Application" and click on "Tools menu" in the menu bar.

4) Using the drop-down menu, go to the library "manager Package Manager Console."

5) Starting from the bottom, "Application," the "Package Manager Console" opened and the cursor blinking.

6) The copied code should be pasted from the internet site through the "help of step 2" using hotkeys "Ctrl and V."

7) Press enter and the application will install automatically.

*Corresponding Author

C. Steps To Load DOM Using HTML Agility Pack

1) Add a DLL reference by going into the “Visual Studio Application” and press on the “Solution Explorer” positioned in the sidebar.

2) Right-click on and then click on “Add Reference,” in the context menu.

3) From the “Reference Manager window,” click on the “browser button” and move to “HAP dll” to select it.

4) Press Ok and go back to the code area of the “Visual Studio application” and insert desired code.

5) Inside the Main-Function, write the following code.

HTML Agility Pack will be used to load the HTML Document

```
HtmlWeb web = new HtmlWeb();
HtmlAgilityPack.HtmlDocument doc = new
HtmlAgilityPack.HtmlDocument();
doc = web.Load("http://technologyCrowds.com");
GetMetaInformation(doc,"description");
```

“GetMetaInformation” method definition.

```
static void GetMetaInformation
(HtmlAgilityPack.HtmlDocument htmldoc, string
value)
{
 HtmlNode tcNode =
htmldoc.DocumentNode.SelectSingleNode("//meta
[@name='" + value + "']");
 string fulldescription = string.Empty;
if(tcNode != null)
{
 HtmlAttribute desc;
 desc = tcNode.Attributes["content"];
 Console.ForegroundColor = ConsoleColor.Red;
 Console.Write(desc.Value);
 Console.ReadLine();
 }
}
```

6) For the main function, the user should click on “Start Button” after saving the code and place cursor on the line “doc=web.load” (“https://technologycrowds.com”); and click on “DocumentNode,” then “InnerHtml.”

7) Click on the “search icon” and a new window will pop up. The new window will have all the “DOM” contents which are “HTML” content.

III. Research Methodology

In the first step we have to briefly list the “URL addresses” of the best online real estate ad web sources, then pass all the URLs in “HtmlAgilityPack” to extract the real estate ad data (e.g. property positions, prices and publication date of the ads, etc.) from numerous web sources. In the next step, with the help of linear regression, we will find the average future growth rate of the prices of each real estate property. In the end, with changes in the current average property prices and the estimated average future growth rates, we create a set of standard rules for making decisions about buying a real estate property. Fig. 1 shows the steps of the research methodology.

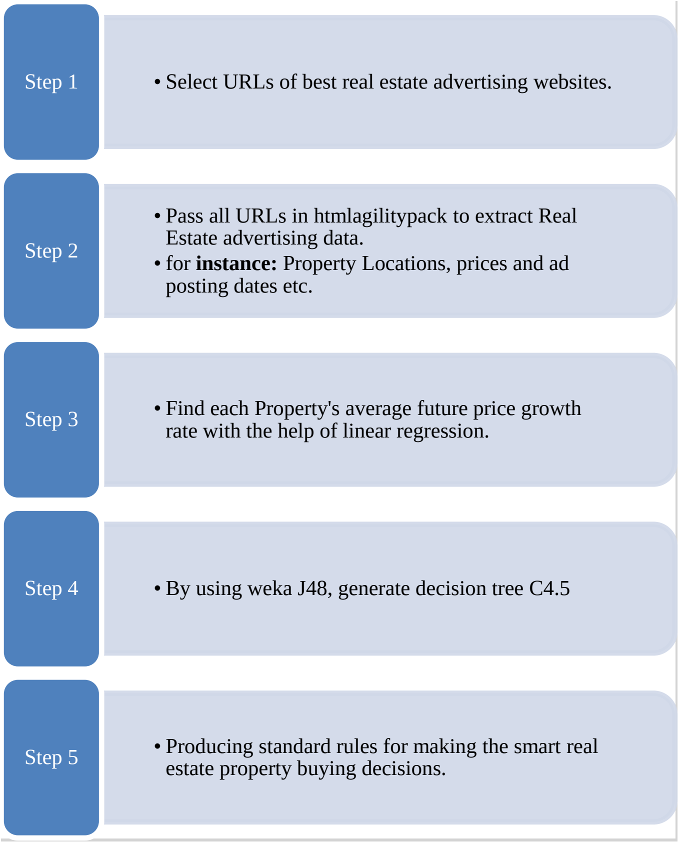

Step 1
• Select URLs of best real estate advertising websites.
Step 2
• Pass all URLs in htmlagilitypack to extract Real Estate advertising data.
• for **instance:** Property Locations, prices and ad posting dates etc.
Step 3
• Find each Property's average future price growth rate with the help of linear regression.
Step 4
• By using weka J48, generate decision tree C4.5
Step 5
• Producing standard rules for making the smart real estate property buying decisions.

Fig. 1. Steps for Executing the Methodology.

IV. Extraction Real Estate Advertisement Data from Various Web Sources

The particular research has used the web-scraping technology i.e. HTML Agility Pack, which uses the Pakistan Online Real Estate websites and their advertisements to bring the desired results. The chosen results are brought by the help of web-scraping technology and Table I shows the average prices of the most popular housing areas in different periods of time.

The future price growth rate of any real estate property has become a very significant factor in order to make real estate property buying decisions. The average prices were seen in the below table i.e. Table I and by the help of those average prices of different time intervals, we can use the linear regression technique to assess the average growth rate of future real estate prices.

TABLE. I. AVERAGE PRICES OF POPULAR LOCATION ON DIFFERENT INTERVALS OF TIME

Sr. No.	**Popular Locations for Houses**	**Avg. Prices From Year 2014 To 2015 (Millions In PKR)**	**Avg. Prices From the Year 2016 To 2017 (Millions In PKR)**	**Avg. Prices From Year 2018 To 2019 (Millions In PKR)**
1	Gulshan-e-Iqbal Karachi	56	65	74
2	Gulistan-e-jauhar Karachi	41	50	58
3	Shahra-e-Faisal Karachi	40	51	59
4	Gulberg Lahore	55	60	65
5	Cantt Lahore	38	40	43
6	Gulberg Islamabad	42	42	44
7	Kashmir Highway Islamabad	100	106	109
8	Lalazar Rawalpindi	7	10	14
9	Gulshan Abad Rawalpindi	9	9	10
10	Saddar Rawalpindi	21	20	21
11	North Karachi Karachi	9	10	15
12	North Nazimabad Karachi	50	54	55
13	Malir Karachi	12	14	15
14	Cantt Karachi	58	59	60
15	Mehmoodabad Karachi	2	4	7
16	Ghauri Town Islamabad	12	13	14
17	Kuri Road Islamabad	20	21	20
18	Bharakahu Islamabad	55	60	70
19	Simly Dam Road Islamabad	89	93	98
20	GT Road Islamabad	7	7	7
21	Harbanspura Lahore	8	8	9
22	Ferozepur Road Lahore	2	3	5
23	Taj Pura Lahore	2	3	6
24	Walton Road Lahore	5	8	11
25	Gulshan-e-Ravi Lahore	25	32	40
26	Misryal Road Rawalpindi	9	10	10
27	Shakrial Rawalpindi	2	3	5
28	Sadiqabad Rawalpindi	6	8	10

V. SIMPLE LINEAR REGRESSION

Simple linear regression establishes the connection between “target variable” and “input variables” by fitting a line, called “regression line” [8]. Generally, the linear equation

$$y = m * x + b \quad (1)$$

The above equation is used to represent the line. Within the equation, “y” acts as the dependent variable, whereas “x” is the independent. “m” depicts the “slope”, and “b” is the “intercept point”.

Machine learning requires the following iteration of the same equation.

$$y(x) = w0+ w1 * x \quad (2)$$

Where “w” denotes the parameters, “x” acts as the input, and “y” is the target variable. Changing values of w0 and w1 will give us different lines, as seen in Fig. 2.

Based on “Linear Regression Analysis” Table II offered the estimated average future property values in the different lengths of time and price growth rate percentage.

As a portion of pre-processing the constant assessed real estate records shown in Table II is renewed to definite form by estimated width of the preferred intervals, as shown in Table III.

Table IV visibly demonstrates the projected real estate property data set that is converted into the categorical form.

Next, the categorical data is given as input to “Decision tree C4.5” (Weka-J4.8)

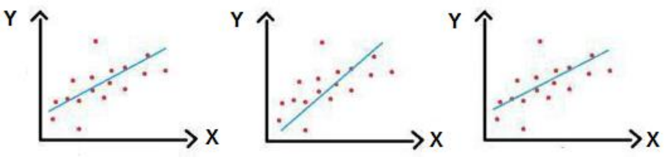


Fig. 2. Linear Regression Lines.

TABLE. II. ESTIMATED AVERAGE FUTURE PROPERTY PRICE GROWTH RATE

Sr. No.	**Popular Locations For Houses**	**Estimated Avg. Future Prices From Year 2020 To 2021 (Millions In PKR)**	**Estimated Avg. Future Prices From Year 2022 To 2023 (Millions In PKR)**	**Estimated Avg. Future Prices From Year 2024 To 2025 (Millions In PKR)**	**Estimated Average Future Price Growth Rate Percentage**
1	Gulshan-e-Iqbal Karachi	81	90	98	10%
2	Gulistan-e-jauhar Karachi	65	73	81	12%
3	Shahra-e-Faisal Karachi	67	76	85	13%
4	Gulberg Lahore	69	74	78	6%
5	Cantt Lahore	45	47	49	4%
6	Gulberg Islamabad	44	45	46	2%
7	Kashmir Highway Islamabad	113	117	121	3%
8	Lalazar Rawalpindi	17	20	23	16%
9	Gulshan Abad Rawalpindi	10	11	11	5%
10	Saddar Rawalpindi	21	21	21	0%
11	North Karachi Karachi	17	20	22	14%
12	North Nazimabad Karachi	58	60	62	3%
13	Malir Karachi	16	18	19	9%
14	Cantt Karachi	61	62	63	2%
15	Mehmoodabad Karachi	9	11	13	20%
16	Ghauri Town Islamabad	15	16	17	6%
17	Kuri Road Islamabad	20	20	20	0%
18	Bharakahu Islamabad	75	82	89	9%
19	Simly Dam Road Islamabad	102	106	110	4%
20	GT Road Islamabad	7	7	7	0%
21	Harbanspura Lahore	9	10	10	6%
22	Ferozepur Road Lahore	6	7	9	23%
23	Taj Pura Lahore	7	9	11	25%
24	Walton Road Lahore	13	16	19	21%
25	Gulshan-e-Ravi Lahore	46	53	60	14%
26	Misryal Road Rawalpindi	11	11	11	0%
27	Shakrial Rawalpindi	6	7	9	23%
28	Sadiqabad Rawalpindi	12	13	15	12%

TABLE. III. CATEGORICAL-PARTITIONING OF ESTIMATED REAL ESTATE PROPERTY DATA SET

Sr. No	**Popular Locations For Houses**	**Partitioned Data**	
		Avg. Current Price Rate Year 2019 (Millions In PKR)	**Estimated Avg. Future Price Growth Rate Percentage**

1	Gulshan-e-Iqbal Karachi	{low, medium, high}
		{ <55,55-60,>60}
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

2	Gulistan-e-jauhar Karachi	{low, medium, high}
		{ <55,55-60,>60 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

3	Shahra-e-Faisal Karachi	{low, medium, high}
		{ <55,55-60,>60 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

4	Gulberg Lahore	{low, medium, high}
		{ <70,70-75,>75 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

5	Cantt Lahore	{low, medium, high}
		{ <40,40-45,>45}
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

6	Gulberg Islamabad	{low, medium, high}
		{ <45,45-50,>50 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

7	Kashmir Highway Islamabad	{low, medium, high}
		{ <110,110-120,>120}
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

8	Lalazar Rawalpindi	{low, medium, high}
		{<10,10-15,>15 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

9	Gulshan Abad Rawalpindi	{low, medium, high}
		{ <5,5-10,>10 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

10	Saddar Rawalpindi	{low, medium, high}
		{ <20,20-25,>25 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

11	North Karachi Karachi	{low, medium, high}
		{ <5,5-10,10> }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

12	North Nazimabad Karachi	{low, medium, high}
		{ <45,45-50,>50 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

13	Malir Karachi	{low, medium, high}
		{ <15,15-20,>20 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

14	Cantt Karachi	{low, medium, high}
		{ <55,55-60,>60 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

15	Mehmoodabad Karachi	{low, medium, high}
		{ <4,4-6>6 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

16	Ghauri Town Islamabad	{low, medium, high}
		{ <5,5-10,>10 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

17	Kuri Road Islamabad	{low, medium, high}
		{ <25,25-30,>30 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

18	Bharakahu Islamabad	{low, medium, high}
		{ <65,65-70,>70}
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

19	Simly Dam Road Islamabad	{low, medium, high}
		{ <70,70-80,>80 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

20	GT Road Islamabad	{low, medium, high}
		{ <5,5-10,>10 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

21	Harbanspura Lahore	{low, medium, high}
		{ <10,10-15,>15 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

22	Ferozepur Road Lahore	{low, medium, high}
		{ <6,6-8,>8 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

23	Taj Pura Lahore	{low, medium, high}
		{ <3,3-5,>5 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

24	Walton Road Lahore	{low, medium, high}
		{ <15,15-20,>20 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

25	Gulshan-e-Ravi Lahore	{low, medium, high}
		{ <25,25-30,>30 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

26	Misryal Road Rawalpindi	{low, medium, high}
		{ <6,6-8,>8 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

27	Shakrial Rawalpindi	{low, medium, high}
		{ <6,6-8,>8 }
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

28	Sadiqabad Rawalpindi	{low, medium, high}
		{ <15,15-20,>20}
			{weak, moderate, strong}
			{ <5%,5-10%,>10% }

TABLE. IV. CATEGORICAL REAL ESTATE PROPERTY DATA SET

Sr. No	Popular Locations For Houses	Avg. Current Price Rate	Estimated Avg. Future Price Growth Rate
1	Gulshan-e-Iqbal Karachi	High	Moderate
2	Gulistan-e-jauhar Karachi	Medium	Strong
3	Shahra-e-Faisal Karachi	Medium	Strong
4	Gulberg Lahore	Low	Moderate
5	Cantt Lahore	Medium	Weak
6	Gulberg Islamabad	Low	Weak
7	Kashmir Highway Islamabad	Low	Weak
8	Lalazar Rawalpindi	Medium	Strong
9	Gulshan Abad Rawalpindi	Medium	Moderate
10	Saddar Rawalpindi	Medium	Weak
11	North Karachi Karachi	High	Strong
12	North Nazimabad Karachi	High	Weak
13	Malir Karachi	Medium	Moderate
14	Cantt Karachi	Medium	Weak
15	Mehmoodabad Karachi	High	Strong
16	Ghauri Town Islamabad	High	Moderate
17	Kuri Road Islamabad	Low	Weak
18	Bharakahu Islamabad	Medium	Moderate
19	Simly Dam Road Islamabad	High	Weak
20	GT Road Islamabad	Medium	Weak
21	Harbanspura Lahore	Low	Moderate
22	Ferozepur Road Lahore	Low	Strong
23	Taj Pura Lahore	High	Strong
24	Walton Road Lahore	Low	Strong
25	Gulshan-e-Ravi Lahore	High	Strong
26	Misryal Road Rawalpindi	High	Weak
27	Shakrial Rawalpindi	Low	Strong
28	Sadiqabad Rawalpindi	Low	Strong

VI. DECISION TREE C4.5

Decision tree refers to a “supervised classification method” that is a structure in which the non-terminal nodes indicate the test of one or more features, and the terminal nodes indicate the result of the decision [9]. It has been apprehended from the studies that the basic algorithm for determining the tree ID3 derivation has been enhanced by the C4.5 algorithm [10]. The unique C4.5 version called J4.8 has a WEKA classification package [11]. In C4.5, the information gain ratio and its measurements are used as a splitting principle, respectively [12]. The steps of this algorithm are given as follows:

Step 1: The set ‘t’ is a set of class labels for tuple training. If an output test is selected, the sample ‘t’ training set must be split into subsets {T1, T2...Tn}. So, the entropy of the set-T can be calculated (in bits).

$$info(T) = -\sum_{i=1}^{k}((freq(C_i,T)/|T|) \times \log_2(freq(C_i,T)/|T|)) \quad (3)$$

Step 2: Divide the training sample by the value of the specified attribute, by which the value of property T will be:

$$infox(T) = -\sum_{i=1}^{n}((|T_i|/|T|) \times info(T_i)) \quad (4)$$

Step 3: Afterwards, the difference between basic information requirements and new information is referred by the information gain. The equation (3) and equation (4) can provide a gain standard:

$$Gain(X) = info(T) - infox(T) \quad (5)$$

Step 4: When building a dense decision tree, the quality of the gain is beneficial, but the test has significant disadvantages because many outputs have large deviations. Therefore, it has to be determined by standardization:

$$Split - info(X) = -\sum_{i=1}^{n}((|T_i|/|T|)\log_2(|T_i|/|T|)) \quad (6)$$

The new gain standard is represented as:

$$Gain - ratio(X) = gain(X)/split - info(x) \quad (7)$$

The Real Estate training dataset (see Table V) is provided as input to “WekaJ48”.

TABLE. V. REAL ESTATE TRAINING DATA SET

Sr. No	Popular Locations For Houses	Avg. Current Price Rate	Estimated Avg. Future Price Growth Rate	Class
1	Gulshan-e-Iqbal Karachi	High	Moderate	No
2	Gulistan-e-jauhar Karachi	Medium	Strong	Yes
3	Shahra-e-Faisal Karachi	Medium	Strong	Yes
4	Gulberg Lahore	Low	Moderate	Yes
5	Cantt Lahore	Medium	Weak	No
6	Gulberg Islamabad	Low	Weak	No
7	Kashmir Highway Islamabad	Low	Weak	No
8	Lalazar Rawalpindi	Medium	Strong	Yes
9	Gulshan Abad Rawalpindi	Medium	Moderate	No
10	Saddar Rawalpindi	Medium	Weak	No
11	North Karachi Karachi	High	Strong	No
12	North Nazimabad Karachi	High	Weak	No
13	Malir Karachi	Medium	Moderate	No
14	Cantt Karachi	Medium	Weak	No
15	Mehmoodabad Karachi	High	Strong	No
16	Ghauri Town Islamabad	High	Moderate	No
17	Kuri Road Islamabad	Low	Weak	No
18	Bharakahu Islamabad	Medium	Moderate	No
19	Simly Dam Road Islamabad	High	Weak	No
20	GT Road Islamabad	Medium	Weak	No
21	Harbanspura Lahore	Low	Moderate	Yes
22	Ferozepur Road Lahore	Low	Strong	Yes
23	Taj Pura Lahore	High	Strong	No
24	Walton Road Lahore	Low	Strong	Yes
25	Gulshan-e-Ravi Lahore	High	Strong	No
26	Misryal Road Rawalpindi	High	Weak	No
27	Shakrial Rawalpindi	Low	Strong	Yes
28	Sadiqabad Rawalpindi	Low	Strong	Yes

VII. STAGES TO CREATE "C4.5 DECISION TREE" IN "WEKA J48"

1) Generate datasets in "MS Excel," "MS Access" and save in "CSV" format.

2) Start the "weka Explorer."

3) Open ".CSV" file and change format to "ARFF."

VIII. RULE PRODUCTION USING DECISION-TREE [14,15] FOR MAKING SMART REAL ESTATE PROPERTY BUYING DECISIONS

The corresponding rules are:

R1: IF (Estimated Avg. Future Price Growth Rate=Weak) Then Purchase = No

R2: IF (Estimated Avg. Future Price Growth Rate=Strong) AND (Avg. Current Price Rate=Low) THEN Purchase = Yes

R3: IF (Estimated Avg. Future Price Growth Rate=Strong) AND (Avg. Current Price Rate=Medium) THEN Purchase = Yes

R4: IF (Estimated Avg. Future Price Growth Rate=Strong) AND (Avg. Current Price Rate=High) THEN Purchase = No

R5: IF (Estimated Avg. Future Price Growth Rate=Moderate) AND (Avg. Current Price Rate=Low) THEN Purchase = Yes

R6: IF (Estimated Avg. Future Price Growth Rate=Moderate) AND (Avg. Current Price Rate=Medium) THEN Purchase = No

R7: IF (Estimated Avg. Future Price Growth Rate=Moderate) AND (Avg. Current Price Rate=High) THEN Purchase = No

1) Click "classify tab," then select "J48."

2) Select any suitable test possibility.

3) Click "Start".

Click on "Visualize Tree" option to view the graphical representation of the tree from the pop-up menu. Fig. 3 depicts the graphical form of Weka J48 generated tree [13].

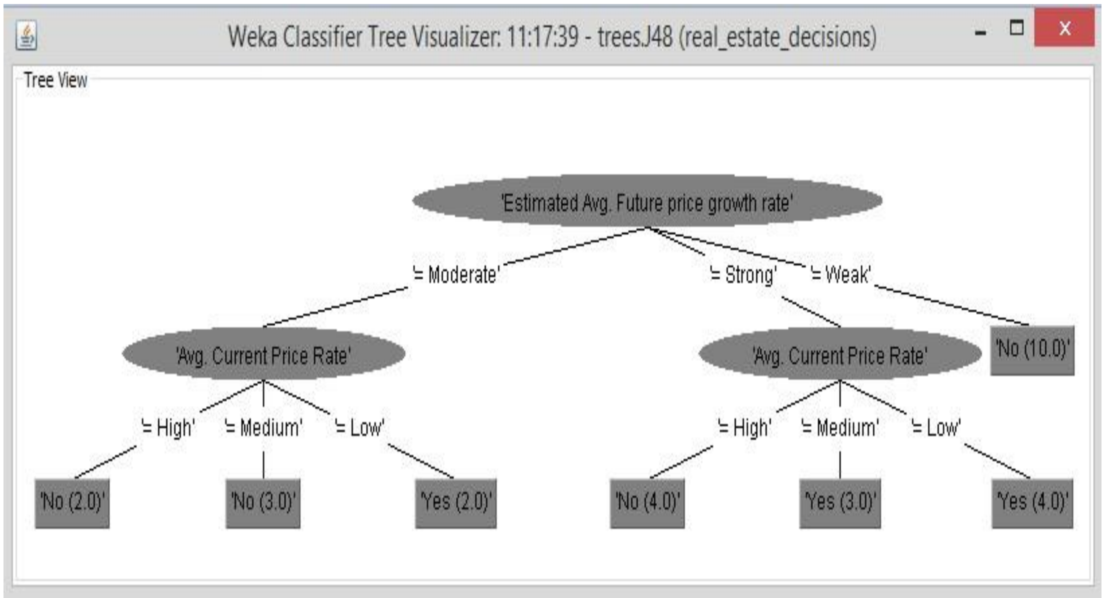

Weka Classifier Tree Visualizer: 11:17:39 - trees.J48 (real_estate_decisions)
Tree View
'Estimated Avg. Future price growth rate'
= 'Moderate'
= 'Strong'
= 'Weak'
'Avg. Current Price Rate'
'Avg. Current Price Rate'
'No (10.0)'
= 'High'
= 'Medium'
= 'Low'
= 'High'
= 'Medium'
= 'Low'
'No (2.0)'
'No (3.0)'
'Yes (2.0)'
'No (4.0)'
'Yes (3.0)'
'Yes (4.0)'

Fig. 3. Real Estate Property Buying Decision Tree C4.5 using Weka J48.

These rules are classified into two classes “YES” and “NO”. The following study discloses only one of the decision-rule for each class.

A. *“NO” Class Rule*

R1: IF (Estimated Avg. Future Price Growth Rate=Weak) Then Purchase = No

It specifies that when the Estimated Avg. Future Price Growth Rate=Weak then purchasing of real estate property is not advisable. Furthermore, 10 training examples support the rule.

B. *“YES” Class Rule*

R2: IF (Estimated Avg. Future Price Growth Rate=Strong) AND (Avg. Current Price Rate=Low) THEN Purchase = Yes

It specifies that when the Estimated Avg. Future Price Growth Rate=Strong and Avg. Current Price Rate=Low then purchasing of real estate property is beneficial. Furthermore, 4 training examples support the rule.

IX. Conclusion

The decision to buy real estate is a substantial financial decision. Buyers should spend a lot of time choosing the best property to buy from all available options. This research concludes that there are no existing standard rules for making smart real estate purchase decisions. However, we propose a method that can generate standard rules for selecting the best real estate property to buy through web scraping technology and machine learning algorithms. This research will save buyers’ time and provide a complete guide to make smart real estate buying decisions.

Acknowledgment

Authors would like to thank Center of Innovations in Computer Science (CICS), Sir Syed University of Engineering and Technology for providing resources and support to perform experiments. This work is also supported by the University of Nottingham, UK.

References

[1] Brueggeman, William B., and Jeffrey D. Fisher. Real estate finance and investments. New York, NY: McGraw-Hill Irwin, 2011.

[2] Vargiu, Eloisa, and Mirko Urru. "Exploiting web scraping in a collaborative filtering-based approach to web advertising." Artif. Intell. Research 2.1 (2013): 44-54.

[3] Hssina, Badr, et al. "A comparative study of decision tree ID3 and C4. 5." International Journal of Advanced Computer Science and Applications 4.2 (2014): 13-19.

[4] Wijaya, James. "Ekstraksi Teks Pada Halaman Website Renungan Rohani Menggunakan HTML Agility Pack." (2019).

[5] https://html-agility-pack.net/

[6] Uzun, Erdinç, et al. "Evaluation of Hap, AngleSharp and HtmlDocument in web content extraction." International Scientific Conference’2017 (UNITECH’17). 2017.

[7] Álvarez-Sabucedo, L. M., Luis E. Anido-Rifón, and Juan M. Santos-Gago. "Reusing web contents: a DOM approach." Software: Practice and Experience 39.3 (2009): 299-314.

[8] Bangdiwala, Shrikant I. "Regression: simple linear." International journal of injury control and safety promotion 25.1 (2018): 113-115.

[9] Siahaan, Hasudungan, et al. "Application of Classification Method C4. 5 on Selection of Exemplary Teachers." Journal of Physics: Conference Series. Vol. 1235. No. 1. IOP Publishing, 2019.

[10] Sathyadevan, Shiju, and Remya R. Nair. "Comparative analysis of decision tree algorithms: ID3, C4. 5 and random forest." Computational intelligence in data mining-volume 1. Springer, New Delhi, 2015. 549-562.

[11] Bhargava, Neeraj, et al. "Decision tree analysis on j48 algorithm for data mining." Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering 3.6 (2013).

[12] Quinlan, J. Ross.C4. 5: programs for machine learning. Elsevier, 2014

[13] Drazin, Sam, and Matt Montag. "Decision tree analysis using weka." Machine Learning-Project II, University of Miami (2012): 1-3.

[14] Jain, Rajni. "Rule generation using decision trees." *IASRI* (2012).

[15] Al-Radaideh, Qasem A., Emad M. Al-Shawakfa, and Mustafa I. Al-Najjar. "Mining student data using decision trees." International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan. 2006.

TADOC : Tool for Automated Detection of Oral Cancer

Dr. Khalid Nazim Abdul Sattar
Department of CSI, College of Science
Majmaah University, Majmaah 11952, Saudi Arabia
ORCID: 0000-0002-0759-0512

Abstract—**Cancer is a group of related diseases and it is necessary to classify the type and its impact. In this paper an automated learning-based system for detection of oral cancer from Whole Slide Images (WSI) has been designed. The main challenges of the system were to handle the huge dataset and to train the machine learning model as it consumed more time for each iteration involved. This further increased the time consumed to get a proper model and decrease of freedom for experimentation. Other important key features of the system were to implement a futuristic deep learning architecture to classify small patches from the large whole slide images and use of carefully designed post-processing methods for the slide-based classification**.

***Keywords**—Cancer; CT Scan; MRI Scan; Machine Learning; Deep learning; Convolutional Neural Network (CNN); Whole Slide Image (WSI); Residual Networks (ResNets)*

I. Introduction

The signs and symptoms of cancer are not visible initially, they are only visible when the mass grows, the growth of abnormal cells in the human body results in different types of cancer affecting the surrounding tissues. Tumors result from the growth of the extra cells which divide without stopping. The advancement of these tumors via the blood or lymph system of the humans result in new tumors away from the place of origin [2]. Several forms of presence of cancer are, limitless number of cell division, promotion of blood vessel construction and avoidance of programmed cell death. Survey based on several factors such as lifestyle, environment, inherited genetics shows that the death rate of patients with cancer are more prone to suicide when compared to the normal people.

A. Background Study

The nomenclature for various types of cancers are usually based on the organs or tissues from where the cancers origin. Doctors use a combination of tests to diagnose the existence of cancer cells in the body. Cancers that comprise under Head and Neck cancer are Lip or Oral cavity cancer, Mouth cancer, Oral cancer, etc. [2].

The cancers are categorized into major five types based on the type of association with the cell they originate from namely, carcinoma, sarcoma, leukemia, lymphoma and myeloma, brain tumors [1].

Some of the often-used tests for identifying the cancerous cells include the following:

1) Laboratory testing: This is a very primary method to determine cancer which can help to rule out other diagnostic procedures.

2) Biopsy: This type of test involves taking a sample of the tissue from a cancerous lesion and to subject it to further laboratory procedures.

3) CT scan: An advanced technique to the regular X-ray method that helps the doctor to scan more details.

4) MRI scan: A technique that makes use of magnets, radio waves and a computer to provide detailed analysis [3].

B. Types of Treatment

A lot of research has been done towards the treatment procedures referring to the types of cancers detected. The different types of treatment methods are as listed below listed with their scope of treatment [2].

- Surgery: follows a procedure-based method to treat the cancer.
- Radiation Therapy: radiation of high dosage is used to kill the cancer cells and to reduce / shrink the tumors.
- Chemotherapy: use of drugs.
- Targeted Therapy: in this type of treatment, the cancer cells can grow, divide and spread.
- Stem Cell transplant: in the process of chemotherapy or radiation therapy , the patient under treatment suffers loss of blood and this is considered as a supplement technique to restore blood forming stem cells.
- Precision medicine: in this type the doctors diagnose and treat the patients based on the genetic history [2].

C. Problems in Manual Diagnosis

a) Delay in diagnosis is the main issue with the manual diagnosis of cancer. It involves extremely skilled labors and the, number of diagnosis tests being requested is growing exponentially.

b) Hinders the possibility of early recognition of tumor grade due to the above stated problems of time consumption for proper diagnosis.

c) Obstruct the provision of instant diagnosis report as the conclusions should be drawn out carefully without causing any fatalities to occur.

d) Enormous work strain of pathologists is a real concern, and this also drops down the accuracy of the pathologist's prediction.

The designed system aims to achieve the following:

1) To handle WSI(Whole Slide Image) i.e. to find an effective way to open the Whole Slide Images instead of opening it in a document viewer with multiple levels of image visible irrespective of its relevance.

2) To create patches from WSI and train the Deep Learning model for prediction.

3) To train the system from the patches generated as mentioned in the above step, such that the trained model will have appropriate weights attached to each parameter after analyzing thousands of patches.

4) To predict tumorous regions and generate the heat map from the prediction model, this will highlight the regions which have high probability of cancer.

The rest of this paper is organized as follows. Section II proceeds with research background and methodology. In Section III, a brief description about deep learning and the details of each part of the implementation is discussed. The result outcomes are as shown in Section IV. Finally, the conclusion and future work are followed in Sections V and VI.

II. Research Background and Methodolody

With the aid of Machine learning, the problems based on appropriate data that will fit into the designed models by using different learning algorithms has been discussed [6][7][26].

Kumar et.al have proposed detection of cancer via microscopic biopsy images through a set of features which were interpretable.

M. Praveen Kiruba bai [54] has explained about the different consequences and techniques related to the detection of Oral Cancer and it has been observed that the Oral cancer on detection at early stage is curable.

Komura, D. and Ishikawa S [25] have explained the techniques for histopathological image analysis using machine learning, the authors have also discussed the importance of collaborating WSIs data based on common criteria.

Since images comprise of several overlapping objects and clusters, an automated system for detecting and classifying the microscopic biopsy images has been proposed in [55].

The importance and applications of medical image analysis using deep learning method has been discussed in [28][50][56].

A. *Conventional Techniques*

A digital image is a collection of a Pixel or Pel refers to the finite number of elements which are specific to their value in a digital Image [7].

The Image processing involves a 3-stage process namely importing by using image acquisition tools, analyzing and manipulating the image followed by generating a report based on the features of interest.

Digital image processing covers several areas of importance such as in the field of medicine, pattern recognition, video processing, image sharpening and restoration [5].

Preprocessing is considered to be one of the elementary steps in image research, which will ease the user to make the image representation in such a way that the application of algorithms will be much easier for various other operations such as segmentation, feature extraction and so on [11][12][19]. The separation of foreground from the background is a vital part for image processing and computer vision as it reduces the computational resources utilized [13].

Fig. 1 illustrates the different methods as used in image preprocessing.

- Histogram Equalization: This method makes use of the cumulative distribution function associated with the image which is the sum of all probabilities of the image in its domain [12][14][15][16]. In histogram equalization the images are processed by modifying the intensity distribution of the histogram associated with the image.

- Mean Filter: an easy method to diminish the noise in the image, that considers of removing pixel values which are misleading of its surrounding value by replacing them with the mean value of its neighbors.

- Median Filter: In this the median values replace the neighboring pixel values [14][17].

B. *Image Segmentation*

The area of interest through different methods from an image viz. cell, nuclei or tumor can be obtained by Image segmentation [18].

The various Image Segmentation methods are as illustrated in Fig. 2 which aid in the diagnosis [19][20].

C. *Feature Extraction*

The prime focus of this method is to detect, isolate distinct portions and features of images. The features extracted are then fed into machine learning algorithm for classification [23][24].

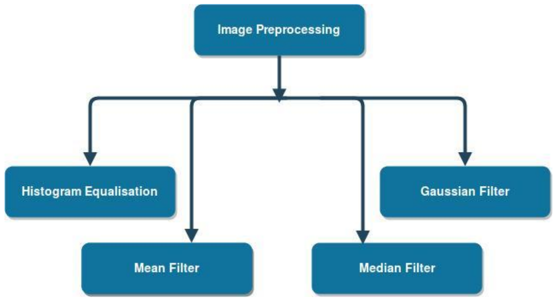


Fig. 1. Image Preprocessing Methods.

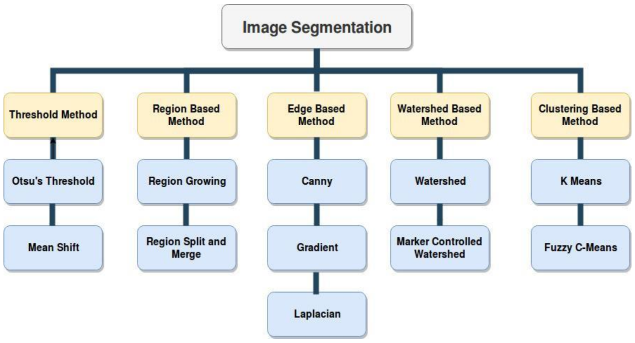


Fig. 2. Image Segmentation Methods.

D. Classification

With the completion of Feature extraction stage, the output which is in the mathematical form is fed into the machine learning algorithm. The taxonomy of classification algorithms for classification purposes, is as shown in Fig. 3.

E. Disadvantages of Conventional methods

- Time Consuming and processing of image takes very long time.
- Choosing appropriate method for each step for processing images.
- Choosing region of interest and appropriate segmentation method.
- Developing proper feature extraction algorithms. Without proper feature extraction the training model and the prediction accuracy will be improper.
- Choosing an appropriate classification algorithm for classification of an image based on the features extracted.

Conventional glass slides are scanned to create digital slides which is referred as Whole Slide Images (WSI). These images have gained beneficiary results in field of education, diagnosis, research. WSI has avoided variance of slide quality by reproducing the same image with the exact orientation. Due to its high image resolution WSI has provided an opportunity of feasible diagnosis for research [29]. A digital WSI is represented as a pyramid with different magnification levels.

For computing resources such as processing power, advanced software is easily available now, digital images have gained wide variety of applications in pathology [30].

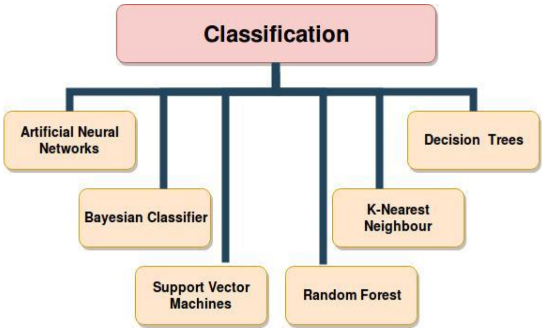


Fig. 3. Taxonomy of Classification Algorithms.

Fig. 4. Whole Slide Images with different Magnification Levels.

There are many challenges that must be addressed while utilizing the WSI, this is because each WSI will occupy large storage space due to its high resolution. Hence storage, transmission and interoperability of WSI are challenging tasks. WSI acquired from different microscopic instruments may have different resolutions and scales of magnification as shown in Fig. 4. The format specification of WSI is not universal which leads to a conflict in viewing, analyzing, accessing with software [31]. Even though WSI enables easy processing facilities of pathological images, these are some of the complexities in handling those images [32].

F. Patch Generation

Patches are sub-images derived from the original image as shown in Fig. 5. Patch can be uniquely identified by horizontal and vertical location inside image, coordinate of center of patch and its size. Patches can be extracted by calculating pixel location of the square when the location and the size are specified. Global features contribute to extraction of texture information, color distribution or whole image information. Information accessed from the global features often turn out to be inadequate, whereas local features like patches will suit to represent restricted region of complex images.

Extraction of these patches can be done through various methods.

- Grids point specification

Regular grid of desired patch size is projected on the image which provides the points to extract. Gaps might be included between the patches depending whether they overlap or not.

- Random point specification

This is like grid point specification except that this chooses the points in random. Hence this is distributed over the image.

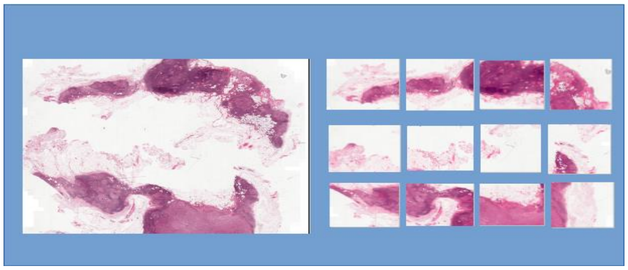


Fig. 5. Patches from WSI Image.

- Interest point specification

Region of Interest is focused and the points inside the same is considered for generating the patches.

G. Advantages of patch-based Approach

- Recognition of the object is location independent. Object that must be recognized might be present in different location in different images [49]. As it is patch based approach, object can be identified irrespective of the location.
- Identification of partial part of the object. Patch based approach helps to identify the objects present in the image even if it is partially occluded.
- Irrespective of size scaling in different images, object can be identified depending on the patch size [33].

Patch based CNN was specifically used in the Music score images [34]. CNN used in the proposed system consists of three convolutional layers which takes in the input. Output of these layers are fed into max pooling and LRN layers. Three fully connected layers consists of 512 neurons each.

This model also consisted of two dropout layers and they were termed as dense1 and dense2 probability of 50% drop probability. Glorot Initialization and ReLU activation are used for initialization and activation respectively for convolutional fully connected layers.

According to the paper patch-based CNN approach has provided promising results in solving writer classification problems.

A Patch Strategy for Deep Face Recognition [35],proposes a system that would take online cropped images as input for face recognition. Multibranched CNN that learn from each patch and entire face representation is done by considering all the patches is used. AlexNet and ResNet pre-trained CNN models are used for analyzing the efficiency of the method. As an end to end training model, usage of both global and local features is done effectively. Six patches of size 136x136 pixel with facial key points from aligned face images are considered.

These patches are passed onto pooling and convolutional layers. Feature fusion is accomplished by fully connected layers. This method boosts the performance of face recognition as it enhances the representation of local features.

In [36] the researchers propose a system with multiscale version of the patches as input. Down sampling is carried by decimating smooth version and up sampling is carried by nearest neighbor interpolation. The proposed system yields smooth and compact segmentation results.

Comparison of Deep Learning patch-based frameworks such as ConvNet, AlexNet and VGG models was carried by training and testing these models with publicly available, high resolution datasets. Varied patch dimension such as 11x11, 21x21, 29x29, 33x33, 45x45 are considered for comparing the accuracy rates and to choose the appropriate patch size for the model [37]. Small patch size turned out to affect the quality and robustness of features in deep layers.

Authors propose patch based Deep Learning approach to explore subtypes of cancer [27][38]. Even though CNN has acquired prominence in image classification, handling high resolution image implies high computational cost. Training CNN directly with Whole Slide Image (WSI) of size merely gigabytes would lead to down sampling and data inefficiency. Hence patch based CNN model for lung cancer subtype classification was proposed by Le Hou.

H. CNN Architectures

CNN takes input in the form of a bunch of arrays, the data is readily available in the form of images and follows the deep feed forward mechanism of network. Images are multi-dimensional array with each unit holding the pixel values and intensities.

CNNs are multi-layered neural networks which can be further subdivided into convolutional and pooling layers. The Fig. 6 as shown below illustrates the representation of a CNN.

The development of a system is based on how neurons work and therefore from the human brain itself. Numerous applications like document processing, semantic analysis of documents, sounds and images have been created using the CNNs already . The document processing system uses a CNN and can as well be trained to implement constraints on languages [39][40][41][42].

There is another variant of the CNNs called as the Fully Convolutional Network (FCN) widely being used for some of the above stated cases as shown in Fig. 7. As already discussed above, CNN will have multiple layers and each layer is a 3D array, where 2 of the three dimensions are spatial dimensions and the other feature is the feature dimension. If the representation of 3 layers is x * y * z, then the first layer i.e. x * y is also the image dimension in pixels.

The efficiency of a Deep Neural Network can be intensified by boosting the depth and its width (size of the network). The easiest way of acquiring models with higher accuracy for gigantic amount of data can be achieved by intensifying the depth and width of the network. But this method has a major drawback in the form of the amount of input features that would be dealing with, which certainly leads to overfitting [43].

I. Residual Networks (ResNets)

In Residual Networks (ResNets), the neural network is broken into small pieces and link the pieces through skip or shortcut type of connections that will form a big network.

Based on the type of input and output dimensions, the residual networks takes into account 2 types of blocks namely the identity block where the input and output activations are similar, while in the convolution block of connection the dimensions differ, as shown in Fig. 8, 9 below depict the representation of these 2 types ResNets blocks.

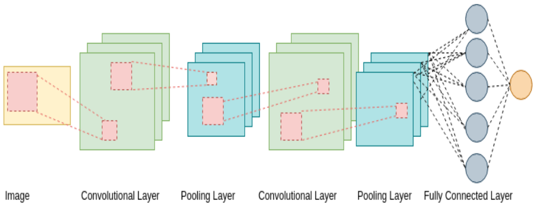


Fig. 6. Representation of Convolutional Neural Network.

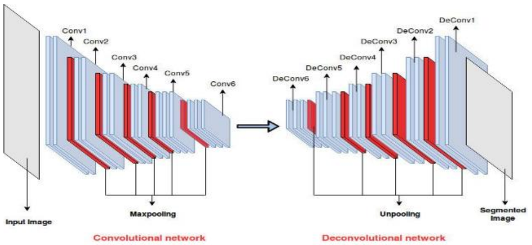


Fig. 7. Representation of Fully Convolutional Neural Network.

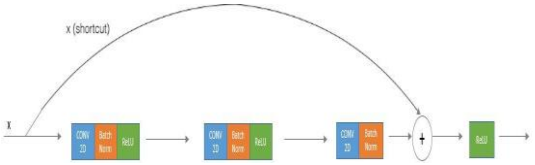


Fig. 8. Representation of Residual Neural Networks Identity Block.

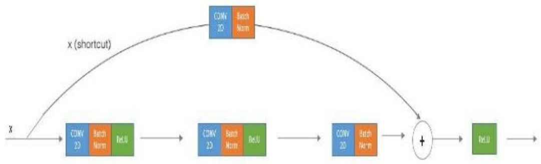


Fig. 9. Representation of Residual Neural Network.

III. Implementation

A. System Design

The data set contained images from a disparate patient population who had oral cancer [46]. The image quality also plays a great role as we can get a better prediction model with an image with a higher resolution [53]. Sometimes the model of image acquisition will have unnecessary variation unrelated to classification levels [44][45].

When the image is fed to the system in a correct format, it undergoes processing through different modules as shown in the data flow diagram in Fig. 10.

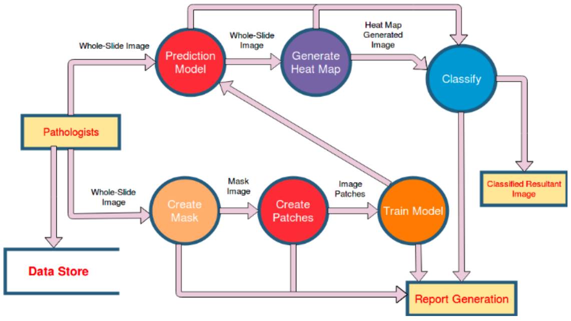


Fig. 10. Proposed Design.

B. Model Training in Deep Learning

A machine learning algorithm has been devised for the model used with the following steps during the training process:

[Step 1]: Define Appropriately the Problem (objective, desired outputs).

[Step 2]: Gathering/ Collection of data.

[Step 3]: Set up an evaluation protocol.

[Step 4]: Formulate the data (viz missing values, Categorical values).

[Step 5]: Split the data appropriately.

[Step 6]: Generalize between overfitting and underfitting problems.

[Step 7]: Summarize the learning process of a model.

[Step 8]: Develop a benchmark model.

[Step 9]: Developing a better model & tuning its hyper parameters to get the best performance possible.

IV. Results and Screen Shots

A. Trained Model Results

As per the survey, even though the work started out with a Fully Convolutional Network (FCN), that is being executed with Keras over TensorFlow, the system had to stop to avoid loss of models which were under training, due to power outages. It was found that using PyTorch would give automatic checkpoints for the models under training till the point of failure, that shifted the focus to the same. The trained model in this was stored in the .ckpt format [9][10][52].

After going through similar implementations for WSI images on other cancer type dataset, referring to one of the recent researches that used ResNet for training models for lung cancer, it was found that there is a similarity in the models used based on same image resolutions. The checkpoints obtained when testing the model with the said dataset, was used further in the process for heatmap generation and other further evaluations.

B. Prediction and Heatmap Generation

In order to understand and interpret the trained model in medical image analysis, visualization of the results is important factor. Most of the times, prediction calculation involves mathematical approach to obtain the probability calculated for that dataset by trained model, based on its knowledge gained during training process [47][48]. Such aspect does not provide more clarity or the evidence to trust the trained model. Hence heatmap generation comes into picture.

There are various methods and readily available python modules to carry out this task. Activation functions and optima’s that are chosen during the training process plays important roles while generating heat map. Approach that is opted to obtain the same in this project is like that of window slide probability calculation. That is, probability of each patch generated from test WSI being tumorous is calculated and is stored in NumPy array. Thus, region prone to tumorous are highlighted in the heatmap.

The output obtained from the prediction algorithm, which was in the NumPy format is converted or depicted as an image.

The above figure is a set of sample images that has undergone a heatmap based prediction. The first image i.e. Fig. 11(a) gives us a glimpse of the actual WSI image under consideration. The next image Fig. 11(b) is the label or mask which is the information about the image under prediction. In this case, the image is cancerous, and the white area is marked as cancerous by pathologists. The next three figures illustrate the prediction heat map obtained for the above image using various models for prediction.

The heat map so obtained is a clear indicator of presence or absence of cancer in each slide provided, the model under evaluation is accurate. These heatmaps are particularly useful for pathologists as they mark the area under suspicion and that part of the slide can be easily selected and observed by any pathologist.

C. *User Interface*

The user interface as shown in Fig. 12(a) to 12(e) is a native application developed for ubuntu operating system using an open source software called PyQt5. Even though a web application would have been easily accessible to everyone, it was not considered due to the obvious reasons of data size and bandwidth capacity [4][8][51].

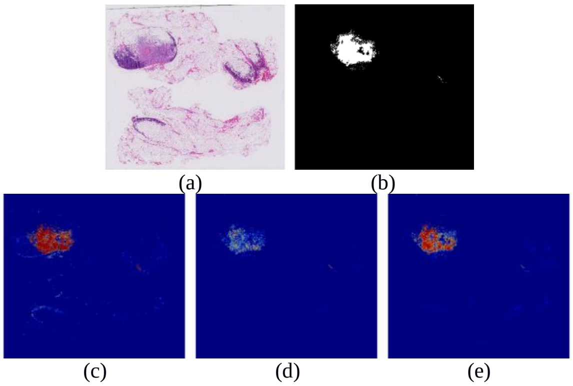
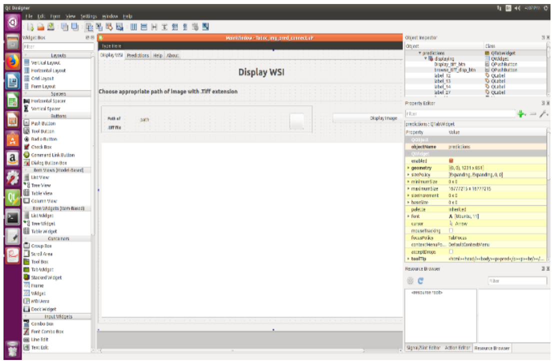
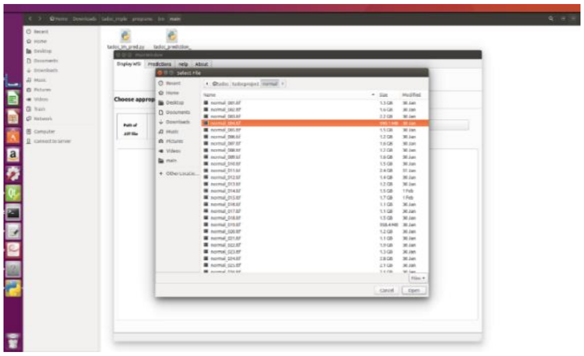


Fig. 11. Stage Wise Depiction of a Heat Map Generation from a Whole Slide Image.

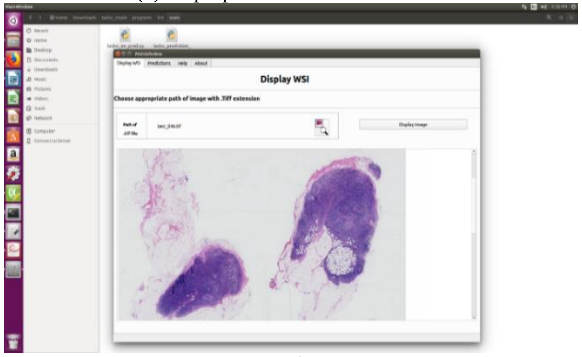


Choose appropriate path of image with .Tiff extension

(a). GUI Designing using the PyQt5 Designer.

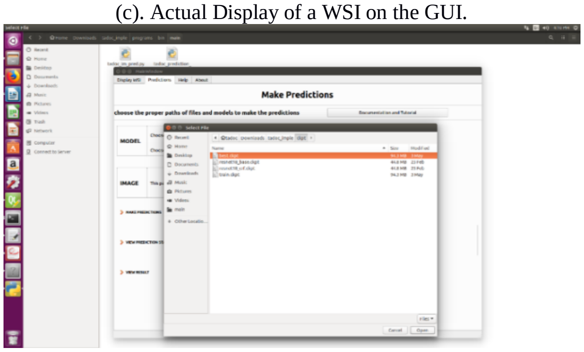


(b). Pop-up Window- File Selection.



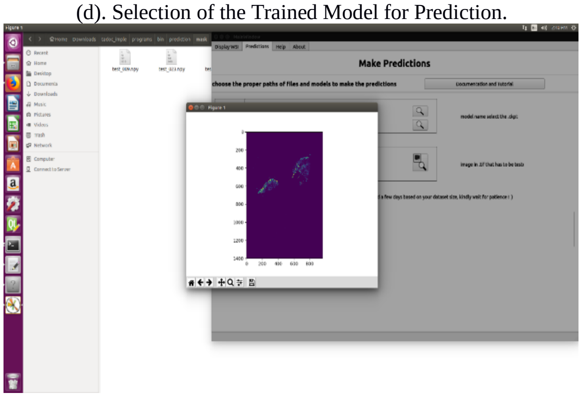
Choose appropriate path of image with .Tiff extension

(c). Actual Display of a WSI on the GUI.



choose the proper paths of files and models to make the predictions

(d). Selection of the Trained Model for Prediction.



choose the proper paths of files and models to make the predictions

Fig. 12. (e). Final Prediction Image in the form of Heat Map.

V. Conclusion and Findings

A learning-based system for automated detection of oral cancer from whole slide images (WSI) has been presented. The main challenges of the system were to handle the dataset as it was huge, to train the machine learning model as it took huge amount of time to get each iteration of the model [6][51]. This further led to the increased time consumed to get a proper model and decrease of freedom for experimentation.

Other important key features of the system were to implement a futuristic deep learning architecture to classify small patches from the large whole slide images and use of carefully designed post-processing methods for the slide-based classification [39]. Classical methods in histopathology were mainly focused on image analysis tasks [20][21][22].

VI. Future Work

The proposed method utilizes ResNet, short for residual neural network deep network architecture. Based on the results of many such experimental results, integrating deep learning-based approaches into clinical practices can bring vast improvements in speed, accuracy, reproducibility, reliability and clinical value of pathological diagnoses.

Acknowledgment

The author would like to thank Dr. Harsha S Jois, Associate Professor, Department of ISE, Jyothy Institute of Technology, Bengaluru, India for his useful critiques and constructive suggestions.

References

[1] Cancer Research UK: www.cancerresearchhuk.org

[2] National Cancer Institute: www.cancer.gov

[3] Mayo Clinic: www.mayoclinic.org

[4] Digital Design Journal: www.digitaldesignjournal.com

[5] Tutorials Point: www.tutorialspoint.com

[6] Digital Ocean: An Introduction to machine learning: www.digitalocean.com

[7] GeeksForGeeks: www.geeksforgeeks.org

[8] GitHub: ASAP- Automated Slide Analysis Platform: www.githhub.com

[9] Facebook Code: www.code.fb.com

[10] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. and Lerer, A, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[11] Anuradha, K. and Sankaranarayanan, K., Detection of oral tumor based on marker-controlled watershed algorithm. International Journal of Computer Applications , 52 (2), 2012.

[12] Rejintal, A. and Aswini, N., “Image processing-based leukemia cancer cell detection,” In 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)(pp. 471-474),2016.

[13] Bhattacharjee, S., Mukherjee, J., Nag, S., Maitra, I.K. and Bandyopadhyay, S.K., “ Review on histopathological slide analysis using digital microscopy,” International Journal of Advanced Science and Technology, 62, pp.65-96, 2014.

[14] Kumar, N. R. and J. U. Kumar, "A spatial mean and median filter for noise removal in digital images," International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 4, No. 1, 246-253.

[15] Al-amri, S.S., Kalyankar, N.V. and Khamitkar, S.D., “Linear and non-linear contrast enhancement image,” International Journal of Computer Science and Network Security, 10 (2), pp.139-143, 2010.

[16] Irshad, H., Veillard, A., Roux, L. and Racoceanu, D., “Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential,” IEEE reviews in biomedical engineering, 7, pp.97-114, 2014.

[17] Senthil Kumar, K., Venkata Lakshmi, K. and Karthikeyan, K.,” Lung Cancer Detection Using Image Segmentation by means of Various Evolutionary Algorithms,” Computational and mathematical methods in medicine, Hindawi,2019.

[18] Belsare D., Mushrif M. M., “Histopathological image analysis using image processing techniques: An overview,” Signal & Image Processing: Int. J. 3(4), 23–26,2012.

[19] Kaur, D. and Kaur, Y., “Various image segmentation techniques: a review,” International Journal of Computer Science and Mobile Computing ,3 (5), pp.809-814, 2014.

[20] Irshad, H., Veillard, A., Roux, L. and Racoceanu, D., “Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential,” IEEE reviews in biomedical engineering, 7, pp.97-114, 2014.

[21] Ali, S. and Madabhushi, A.,” An integrated region-, boundary-, shape-based active contour for multiple object overlap resolution in histological imagery,” IEEE transactions on medical imaging, 31 (7), pp.1448-1460, 2012.

[22] Ali, S. and Madabhushi, A. ”Active contour for overlap resolution using watershed-based initialization (ACOReW): Applications to histopathology,” IEEE International Symposium on Biomedical Imaging: From Nano to Macro (pp. 614-617) 2011.

[23] Amandeep, S. Jain, S. Bhusri, CAD system for non-small cell lung carcinoma using laws‟ mask analysis, international conference on computing for sustainable global development, BVICAM, pp. 6285-6288, 2017.

[24] Di Cataldo, S. and Ficarra, E., “Mining textural knowledge in biological images: Applications, methods and trends,” Computational and structural biotechnology journal, 15, pp.56-67, 2017.

[25] Komura, D. and Ishikawa, S.,” Machine learning methods for histopathological image analysis,” Elsevier, Computational and structural biotechnology journal, 16 , pp.34-42, 2018.

[26] Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V. and Fotiadis, D.I., “Machine learning applications in cancer prognosis and prediction,” Elsevier, Computational and structural biotechnology journal, 13, pp.8-17, 2015.

[27] Nahid, A.A. and Kong, Y.,” Involvement of machine learning for breast cancer image classification: a survey,” Computational and mathematical methods in medicine, 2017.

[28] Hamilton, P.W., Wang, Y. and McCullough, S.J., “Virtual microscopy and digital pathology in training and education,” Apmis, 120 (4), pp.305-315, 2012.

[29] Al-Janabi, S., Huisman, A. and Van Diest, P.J.,” Digital pathology: current status and future perspectives,” Histopathology, 61 (1), pp.1-9, 2012.

[30] Goode, A., Gilbert, B., Harkes, J., Jukic, D. and Satyanarayanan, M.,” OpenSlide: A vendor-neutral software foundation for digital pathology,” Journal of pathology informatics, 4, 2013.

[31] Boyce, B.F., “Whole slide imaging uses and limitations for surgical pathology and teaching,” Biotechnic & Histochemistry, 90 (5), pp.321-330, 2015.

[32] Farahani, N., Parwani, A.V. and Pantanowitz, L.,” Whole slide imaging in pathology: advantages, limitations, and emerging perspectives,” Pathology and Laboratory Medicine International, 7, pp.23-33, 2015.

[33] Hegerath, A., Ney, I.H. and Seidl, T.,” Patch-based object recognition,”(Doctoral dissertation, Diploma thesis, Human Language Technology and Pattern Recognition Group, RWTH Aachen University, Aachen, Germany), 2006.

[34] Hattori, L.T., Gutoski, M., Aquino, N.M.R. and Lopes, H.S., “Patch-Based Convolutional Neural Network for the Writer Classification Problem in Music Score Images,” 2018.

[35] Zhang, Y., Shang, K., Wang, J., Li, N. and Zhang, M.M.,” Patch strategy for deep face recognition,” IET Image Processing, 12 (5), pp.819-825,2018.

[36] Stawiaski, J., “A Multiscale Patch Based Convolutional Network for Brain Tumor Segmentation,” arXiv preprint arXiv:1710.02316, 2017.

[37] Papadomanolaki, M., Vakalopoulou, M. and Karantzalos, K., ”Patch-based deep learning architectures for sparse annotated very high-resolution datasets,”. In 2017 Joint Urban Remote Sensing Event (JURSE) (pp. 1-4). IEEE, 2017.

[38] Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E. and Saltz, J.H.," Patch-based convolutional neural network for whole slide tissue image classification," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2424-2433), 2016.

[39] Maier, A., Syben, C., Lasser, T. and Riess, C.," A gentle introduction to deep learning in medical image processing," Zeitschrift für Medizinische Physik, 2019.

[40] S Tandel, G., Biswas, M., G Kakde, O., Tiwari, A., S Suri, H., Turk, M., Laird, J.R., Asare, C.K., A Ankrah, A., N Khanna, N. and K Madhusudhan, B., "A Review on a Deep Learning Perspective in Brain Cancer Classification," Cancers, 11 (1), p.111, 2019.

[41] Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O. and Hajirasouliha, I.," Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images," EBioMedicine, 27, pp.317-328, 2018.

[42] Choi, H.," Deep learning in nuclear medicine and molecular imaging: current perspectives and future directions," Nuclear medicine and molecular imaging, 52 (2), pp.109-118, 2018.

[43] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A.," Going deeper with convolutions,", In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9), 2015.

[44] The Data Science Blog: www.ujjwalkarn.me

[45] Towards Data Science: www.towardsdatascience.com

[46] The Cancer Imaging Archive (TCIA): www.cancerimagingarchive.net

[47] Reitermanova, Z., "Data splitting," In WDS (Vol. 10, pp. 31-36), Part I, 31–36, 2010.

[48] Diehl, J., "Preprocessing and Visualization," In Seminar Data Mining in WS (pp. 1-21), 2004.

[49] He, K., Zhang, X., Ren, S. and Sun, J.," Deep residual learning for image recognition," In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778), 2016.

[50] Yousoff, S.N.M., Baharin, A. and Abdullah, A., "A review on optimization algorithm for deep learning method in bioinformatics field," In 2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES) (pp. 707-711), 2016.

[51] Syberfeldt, A. and Ekblom, T.," A comparative evaluation of the GPU vs. the CPU for parallelization of evolutionary algorithms through multiple independent runs," International Journal of Computer Science & Information Technology (IJCSIT), 9(3), pp.1-14, 2017.

[52] Wongsuphasawat, K., Smilkov, D., Wexler, J., Wilson, J., Mane, D., Fritz, D., Krishnan, D., Viégas, F.B. and Wattenberg, M., "Visualizing dataflow graphs of deep learning models in TensorFlow," IEEE transactions on visualization and computer graphics, 24(1), pp.1-12, 2017.

[53] Cruz, J. A., and Wishart, D. S. Applications of machine learning in cancer prediction and prognosis. Cancer Inform. 2, 59–77. doi: 10.1177/117693510600200030, 2006.

[54] M.Praveena Kiruba bai," A Survey on the Detection of Oral Cancer," IJIACS,ISSN 2347 – 8616,Volume 6, Issue 7,July 2017.

[55] Kumar, R., Srivastava, R. and Srivastava, S.," Detection and classification of cancer from microscopic biopsy images using clinically significant and biologically interpretable features," Journal of medical engineering, 2015.

[56] A. Qayyum, S. M. Anwar, M. Majid, M. Awais, and M. Alnowami, "Medical image analysis using convolutional neural networks: A review," arXiv preprint arXiv:1709.02250, 2017.

AUTHOR'S PROFILE

Dr. Khalid Nazim Abdul Sattar, Assistant Professor, Department of CSI, College of Science, Majmaah University, Az Zulfi, Kingdom of Saudi Arabia, received his B. E. Degree from Bangalore University, M. Tech from VTU Belagavi, and Ph.D. in Computer Science & Engineering from Singhania University, Rajasthan, India. He has 19 years of research and teaching experience and his research expertise include but are not limited to Image processing, digital signal processing, Artificial Intelligence, Internet of Things, Cryptography, Data & Network Security and Data mining.

Deep Learning based, a New Model for Video Captioning

Elif Güşta Özer[1], İlteber Nur Karapınar[2], Sena Başbuğ[3], Sümeyye Turan[4], Anıl Utku[5], M. Ali Akcayol[6]
Department of Computer Engineering, Faculty of Engineering
Gazi University, Ankara, Turkey

***Abstract*—Visually impaired individuals face many difficulties in their daily lives. In this study, a video captioning system has been developed for visually impaired individuals to analyze the events through real-time images and express them in meaningful sentences. It is aimed to better understand the problems experienced by visually impaired individuals in their daily lives. For this reason, the opinions and suggestions of the disabled individuals within the Altınokta Blind Association (Turkish organization of blind people) have been collected to produce more realistic solutions to their problems. In this study, MSVD which consists of 1970 YouTube clips has been used as training dataset. First, all clips have been muted so that the sounds of the clips have not been used in the sentence extraction process. The CNN and LSTM architectures have been used to create sentence and experimental results have been compared using BLEU 4, ROUGE-L and CIDEr and METEOR.**

Keywords—Video captioning; CNN; LSTM

I. Introduction

In order to facilitate the lives of visually impaired individuals, the new technologies have been developed at last decade. These technologies can help people who are having trouble with their vision ability. About 1.3 billion people have been lived with this problem, in 2018 [1]. The technologies can take these people where they want to go, or they can help them read the texts. In this study, it has been aimed that the visually impaired individuals can perceive the events in their environment. Also, the event analysis over video should be worked in real time. The images detected from the camera are divided into frames in real time. By analyzing these images, the objects are recognized then the differences between the frames are detected and a prediction has been done for the actual action. Then, the basic event in the perceived real-time image is converted into a sentence and finally converted to sound. Thus, visually impaired individuals can be able to learn what is happening around them without the help of another person. Especially individuals who have lost their vision later can easily understand and visualize the descriptions.

In the literature, it is possible to examine the video subtitle creation methods that proposed in this context in two groups. The first of these is generative based methods and the second is retrieval-based methods such as Recurrent Neural Network (RNN). The basic approach of the first group of methods is object recognition on a visual content and the creation of subtitle with natural language creation techniques [2, 3]. The methods in the second group, unlike the first group, utilize from the visual similarities of the visual contents on the data set and the textual similarities of the subtitles, at the same time and select the most likely subtitle for the images from the appropriate subtitles [4, 5].

Yao et al., have developed an automatic video description system [4]. A directory has been created to improve search quality in online videos and to enable visually impaired people to use video definition. Youtube2Text and the DVS dataset have been used as datasets. From the deep learning techniques, combination of RNN and ConvNet models and GoogleNet as the architecture have been utilized. The results have been tested by applying Long Short Term Memory Network (LSTM) model which is a RNN type without soft-attention mechanism and no mechanism. After that the reason for using the soft-attention mechanism, has not been to include in the definition of less important actions and objects in the clip. Experimental results showed that the soft-attention mechanism improves the identification performance.

Venugopalan et al., have developed Sequence-to-Sequence Video to Text (S2VT) model to make description one of the videos there [5]. They implemented the developed model using Microsoft Video Description (MSVD) dataset, MPII Movie Description dataset (MPII-MD) and Montreal Video Annotation Dataset (MVAD) dataset. As the method, LSTM has been used in actualized study. Also, in this study no pooling method has been used. In the proposed model, the successive (sequence to sequence) video frames have been taken as an input and as the output, it has been given successive words. LSTM resolves the video, frame by frame. To do this, convolutional neural network (CNN) output which applied on every frame taken as an input. After reading all frames, the model has been created a sentence from the words. AlexNet and VGG-16 have been used as architecture in this study.

Li et al., have developed an architecture called Residual Attention-based LSTM (Res-ATT) [6]. To describe the video in detail, the mechanism called temporal attention with CNN and LSTM has been applied by them. This mechanism has been used to better identify the important events in the video. They also used a technique they called Residual to avoid the degradation problem when using RNN. MSVD and MSR-VTT datasets have been used to test the architectures which they developed and as the metric, BLUE, METEOR, CIDEr have been used in this study. Also, Microsoft COCO caption evaluation has been used as an evaluation code.

Rohrbach, et al., have studied on HD films [7]. The MPII-MD dataset has been used for this study. AD (Audience Descriptions), defines films for blind or visually impaired people. Multiple sentence definitions and long videos presents in TACoS Multi-Level and YouCook datasets on this topic. However, for shorter term videos, dataset options are increasing. Apart from the definition with AD, the study may also have tasks such as creating a story from relationships in the film and analyzing relationships.

Xu et al., have looked at video captioning from a different perspective [8]. Videos has been modelled as a sequence of frames. The Attentive Multi-Grained Encoder (AMGE) model for the encoder phase has been used.

Krishna et al., have defined the detected events simultaneously with natural language. Using developed model, all events have been identified in a single transition of the video [9]. ActivityNet Captions, has been selected as dataset. A hierarchical RNN structure has been used to provide more detailed events in the videos. Semantic information in videos has been taken as input. This information has been given in the form of an Array structure. Then, LSTM fed. However, there has been time differences between events because detailed events have been described. Events have this time difference have been handled separately. BLEU, METEOR and CIDEr have been used as metric.

Wu et al., have focused on the video classification problem [10]. Short-term events occur on videos and a hybrid model has been created because of these short-term events in the study. Two different feature extraction methods have been used for a given input video: Spatial CNN and Short-term stacked motion optical flows. Extracted features and LSTM models have been fed separately. The results have been combined to make the final prediction. The UCF-101 Human Actions and the Columbia Consumer Videos (CCV) have been selected as the dataset of the study.

Yue-Hei Ng et al., have proposed two deep neural network architectures for combining long-time videos' information [11]. Models has been used to understand the hidden events existing in image in every stream. Various convolutional temporal feature pooling architectures have been tried in the first proposed architecture. They have used LSTM cells connected to the output of the CNN in the second proposed architecture. Performance of the proposed architecture for Sports 1 million dataset is 73.1% and UCF-101 dataset is 88.6%.

Wang et al., have presented a new model that combines audio and visual cues called HACA (Hierarchically Aligned Cross-modal Attentive) network [12]. The proposed new HACA model learns and aligns both global and local contexts between different forms of video. In this study, hierarchical encoder-decoder network including visual attention, audio attention, and decoder attention have been used by them. CNN models that was pre-trained has been used to extract visual features and audio features. For image classification they used the ResNet model and for audio classification. They used the VGGish model. Besides, MSR-VTT has been used for the model training.

In this study, sequence-to-sequence (Seq2Seq) models have been used. The VGG-16 and VGG-19 CNN architectures are used with the LSTM. Developed model has been trained on video to text pairs. In this study, it is aimed that the developed model has been learned to associate a variable-sized square array with a variable-sized word array. The performance of developed model has been evaluated on the Microsoft Video Description Corpus (MSVD) and the BLUE, ROUGE, CIDEr and METEOR metrics.

II. Video Captioning

Video captioning is a popular research field for computer vision, image processing and natural language processing. In video captioning, it is aimed to automatically obtain a natural sentence from a video. However, automatically creating natural language definitions of videos is a challenge for machines. Automatic video description model should be able to express objects and events presented in the video. Automatic video description model also explains their relationships with each other in a natural sentence.

Fig. 1 shows differences between video tagging, image captioning and video captioning. The video tag is the name of a extraction of particular object or event in the video. Image (frame) captioning is automatically generating a single sentence or multiple sentence that define an image. Video captioning should also capture the causality between events, actions and objects, as well as the speed and direction of the objects involved [13].

Video caption generation can be classified in two main categories. Template-based models depend on specific grammar rules. Sequence learning-based model learn the probability distribution of visual content, and create new sentences with syntactic structure. These operations explore general (Seq2Seq) models for the generation of video captions. The sequence learning is shown in Fig. 2. Given an input video, 2D and/ or 3D CNN are used to extract visual characteristics in raw video frames. Mean pooling or soft attention operations are performed on visual features. Then, LSTM is trained.

BLEU [14], ROUGE [15] and CIDEr [16] metrics are used for evaluation of the video captioning task. BLEU is based on precision and only controls the n-gram matches in the estimated and basic references. ROUGE has different n-gram versions and calculates recall. CIDEr measures Term Frequency Inverse Document Frequency (TF-IDF) calculating for each n-gram. The performance of our models are evaluated using these 3 important metrics.

By the increasing interest of the video captioning, several large datasets have been released. The YouTube cooking video dataset (YouCook), contains videos about scenes where people cook various recipes [17]. Similarly, TACoS-ML is mainly built based on MPII Cooking Activities dataset and contains cooking videos and descriptions [18]. M-VAD is composed of about 49,000 DVD movie snippets extracted from 92 DVD movies [19]. MPII-MD is another collection of movie descriptions dataset that is similar to M-VAD. It contains around 68,000 movie snippets from 94 Hollywood movies [20]. We perform all our experiments on the MSVD dataset. Microsoft Video Description (MSVD) dataset [21] is a collection of 1,970 YouTube snippets with human annotated sentences. Comparisons of video captioning datasets have shown in Table I.

TABLE. I. COMPARISON OF VIDEO CAPTIONING DATASETS

Dataset	Context	Source	Video	Clip	Sentence	Word	Duration (hrs)
YouCook	cooking	labeled	88	-	2668	42457	2.3
TACos	cooking	AMT workers	123	7206	18227		-
M-VAD	movie	DVS	92	48986	55905	519933	84.6
MPII-MD	movie	DVS+Script	94	68337	68375	653467	73.6
MSVD	multi-category	AMT workers	-	1970	70028	607339	5.3

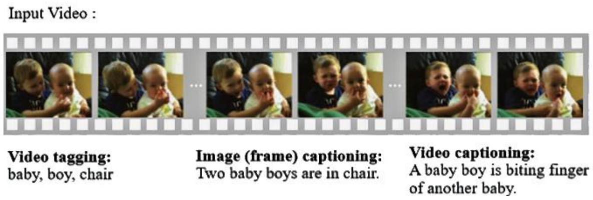

Input Video :
Video tagging:
baby, boy, chair
Image (frame) captioning:
Two baby boys are in chair.
Video captioning:
A baby boy is biting finger of another baby.

Fig. 1. Video Tagging, Image Captioning and Video Captioning.

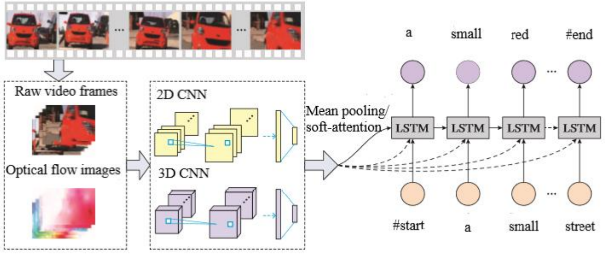


Fig. 2. A Common Architecture with Sequence Learning for Video Captioning.

III. DEEP LEARNING

For years, human being propensity to manage their world more easily by modeling on computers. To overcome this, algorithms that can learn to construct context by themselves have been developed by researchers. Artificial Intelligence (AI) concept has emerged in this development process. AI is a sub-branch of machine learning and uses many non-linear layers for feature extraction.

Neural networks make a trained prediction based on categories and analysis. A machine learning system makes this prediction based on the greatest possibilities. Many of them, learn from their mistakes and this makes them a more accurate system.

The deep learning process is based on learning from data. Computational models consisting of multiple processing layers are used to have a good predictive system in deep learning. The complex structure in complex data sets is discovered using the back propagation algorithm. In this way, more complex concepts are learned using the created hierarchy of concepts. The architectures used in the project are CNN and LSTM.

A. CNN

CNN is an important structure for object detection and classification. The major advantage of CNN is that important features can be detected automatically without any human supervision. To illustrate, when given pictures that are many cats and dogs, the class's unique characteristics are learned for each class.

To give an example structure for the CNN architecture as shown in Fig. 3: first, it takes the regions of the input image one by one under the name of receptive field. Convolution process and pooling are performed on the incoming input in CNN. Feature maps are generated as a result of the convolution layer which is the main block of CNN. Pooling, on the other hand, shortens training time and reduces size to combat overfitting. Property maps are sent to fully connected layer input by making flatten. The classification process is performed on fully connected layers and outputs. The hyper parameters such as the number and convolutional and pooling layers, the number of fully connected layers and activation functions have been determined before training phase.

B. LSTM

The inefficient process of the RNN architecture, which works with backward dependence due to the gradient problem that called vanishing gradient problem, caused this architecture to remain in the background. This problem has been solved by LSTM. LSTM is an effective model for capturing long-term temporal dependencies. It is particularly preferred for speech and text processing.

An LSTM layer consists of a series of blocks that are repeatedly connected, known as memory blocks. Each LSTM layer includes one or more repetitive connected memory cells and input, output and forget gates as shown in Fig. 4. The outputs of the memory cells are connected to all the gates and the cell itself. The gates are optionally a means of transmitting information, and they comprise a layer of sigmoidal neural network and a dot multiplication process. In the sigmoid layer, how much of each component must pass is defined between zero and one. The value of zero means "don't let anything to pass", whereas the value of one means "let everything to pass". The cell condition is like a kind of conveyor belt. Additionally, LSTM has learning rate, hidden layer size and unit parameters. Units are the memory part of LSTM. Increasing the number of units is a widely used option to achieve a powerful model. On the other hand, the training time takes longer time with more unit. This means that the model has learned more.

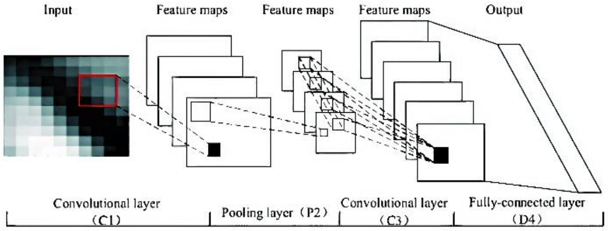


Fig. 3. CNN Structure as an Example.

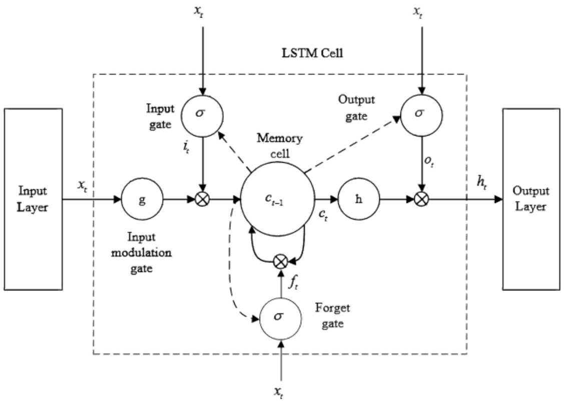


Fig. 4. LSTM Gates.

LSTM, which is used as a decoder within the project, uses feature vectors from CNN for word production and combines language knowledge. LSTM can do the sentence building process because of it can store previously defined objects in its memory. In the word generation phase, the next word is produced using current state and past states. Word generation is continued until end of sentence token is received.

IV. Developed Model

In this study, it has been aimed to develop a system that can accurately express the events in the videos with subtitles. In this section, the structure and steps of the study has been explained. Then, the experimental results of each architecture used in the study have been analyzed.

Fig. 5 illustrates the architecture of the system. The structure of the project consists of feature extraction, learning and prediction steps. In the developed system, videos have been given as input to the system. The videos provided to the system have been converted into frames before being exported to the CNN model to be used and feature extraction was made from these created frames. Caffe has been used for feature extraction in the project.

In the system, Caffe is loaded before feature extraction starts. The models included in Caffe are specially trained models with ImageNet for object detection and UCF101 (Fig. 6) for motion classification. Frames in selected intervals have been selected from the videos. Frames are a series of images. This sequence of images creates the representation of the video. This resized representation sequence has been sent to the selected CNN model as input data. Objects and movements in representations have been determined. Specified properties have been added as attributes to a numpy file. The Caffe has some pre-trained models. In the project, VGG16, VGG19 and HybridCNN models have been used and the results of these models have been examined.

Then, the LSTMs have been used for the second step, the learning process, and the third step, to generate explanations of different lengths. In this step, two LSTM layers have been used as Encoder and Decoder. Using LSTM units in different numbers, the effect on training has been examined. The first layer of LSTM has been used for video processing; the other layer has been used to learn the sentence structure.

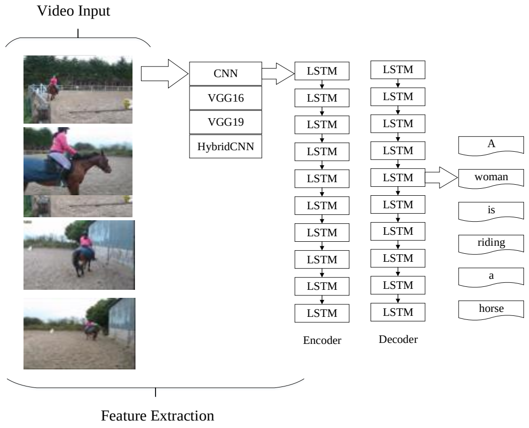


Fig. 5. Implemented Application Model.

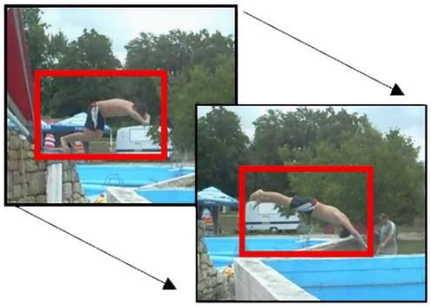
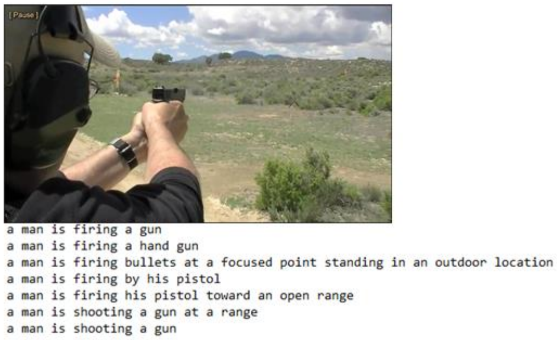


Fig. 6. Detect Motion with UCF101.



```
a man is firing a gun
a man is firing a hand gun
a man is firing bullets at a focused point standing in an outdoor location
a man is firing by his pistol
a man is firing his pistol toward an open range
a man is shooting a gun at a range
a man is shooting a gun
```

Fig. 7. Example of Reference Sentences in the MSVD Dataset.

The properties specified for each video in the feature extraction have been processed in the encoder LSTM layer, creating hidden representation and feeding each other. The purpose of hidden representation is to teach the algorithm its own feature engineering and to make the process more reliable. After this stage is finished for the whole video, the model comes to the decoder layer. The encoder LSTM outputs have been sent to the decoder LSTM units with reference sentences (Fig. 7) that the videos have from the very beginning. Estimates have been compared with actual reference sentences and back propagation is performed in LSTM units.

Loss values are calculated as shown in Fig. 8. It has been ordered in accordance with the sentence structure in English in accordance with the sentence sequence that is learned from the reference sentences.

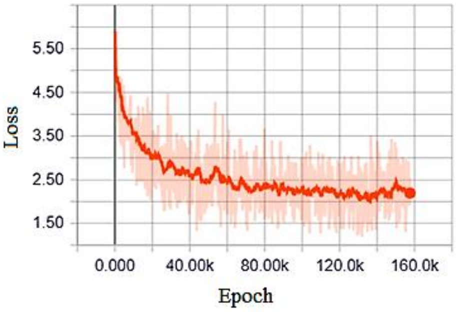


Fig. 8. Change in Loss Value.

During the last step, two LSTM layers have been used as in the training phase. The predicted sentences have been recorded in a file with the name of the video tagging the beginning and end. The purpose of these records is to measure the accuracy of the study. BLEU, CIDEr, ROUGE-L and METEOR metrics have been used for this measurement.

V. Experimental Results

The experimental studies have been conducted using BLEU, ROUGE_L, CIDEr and METEOR. The BLEU algorithm compares the array of expressions generated by the model with the reference expressions of the video and gives scores based on the number of matches. Different n-gram values have been used for BLEU. This means that n expressions have been compared according to the value of n. The difference between ROUGE and BLEU is that BLEU measures the incidence of machine-generated words (and/or n-grams) in the reference summaries. However, ROUGE measures the frequency of words (and/or n-grams) in the machine-generated summaries. CIDEr evaluates the quality of image descriptions. CIDEr measures the consensus between reference sentences and candidate image descriptions. For calculating this metric, each sentence has been represented with a set of 1-4 grams. Finally, METEOR uses harmonic mean of precision and recall of n-gram. It corrects some of the shortcomings of BLEU such as better matching of synonyms, though METEOR and BLEU measures are often used together in evaluation. Co-occurrences of n-grams in the candidate and reference sentences are calculated. Finally, the cosine similarity has been computed between n-grams of the candidate and the references.

The first of the different situations created for the evaluation of the study is the different epoch numbers applied in training. During the training of the properties determined with the VGG16 caffe model, epoch values of 1, 300 and 600 have been applied respectively (Table II). Rapid decrease in epoch score values means that the results obtained from the training are not efficient. The variation of the score between the value of 300 and the value of 600 can be interpreted, as the increase of the epoch value does not have a positive effect on training after a point.

Another situation created for evaluation is the use of different values in LSTM units. The experiment with 1000 epoch has more successful. This is because the back memory is larger. In this evaluation, the difference in CIDEr value is significant. It is seen in Table II that more successful results have been obtained in 300 epochs in experimental studies using 1000 LSTM units. 0.755 BLEU_1, 0.627 BLEU_2, 0.530 BLEU_3, 0.412 BLEU_4, 0.665 ROUGE_L, 0.651 CIDEr and 0.308 METEOR ratios have been obtained.

Finally, the situation evaluated is the results on different models. When the score results have been examined, it has been seen that the models used in feature extraction has a significant effect on the results. It has been seen from the high difference between BLEU and CIDEr that the treatment with VGG19 is more accurate than the other.

The test results in Fig. 9 also explain the difference between CIDEr score values. As a result of the experiments, a number of cases affecting the accuracy of the explanation to be created for a Video were examined and its effect documented with some metrics. Microsoft MSCOCO evaluation scripts have been used to make these comparisons.

TABLE. II. Epoch Test Results with VGG16 + 1000 Unit LSTM

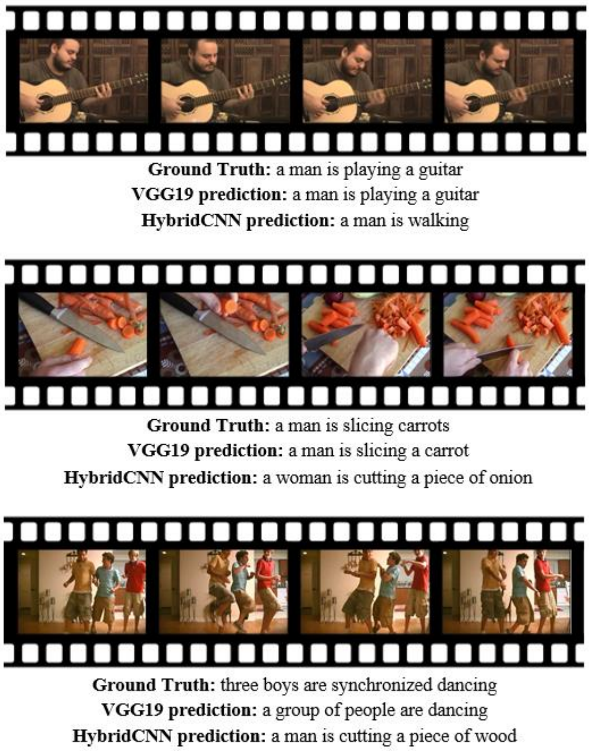
Epoch	**BLEU_1**	**BLEU_2**	**BLEU_3**	**BLEU_4**	**ROUGE_L**	**CIDEr**	**METEOR**
1 epoch	0.568	0.038	0.000	0.000	0.417	0.002	0.081
300 epoch	0.755	0.627	0.530	0.412	0.665	0.651	0.308
600 epoch	0.742	0.596	0.473	0.344	0.656	0.660	0.293

TABLE. III. LSTM Units Test Results with VGG19

Unit	**BLEU_1**	**BLEU_2**	**BLEU_3**	**BLEU_4**	**ROUGE_L**	**CIDEr**	**METEOR**
200 LSTM	0.593	0.436	0.328	0.215	0.575	0.298	0.236
1000 LSTM	0.719	0.569	0.460	0.355	0.639	0.646	0.294

TABLE. IV. Experimental Results for HybridCNN and VGG19 with 1000 Epoch

Model	**BLEU_1**	**BLEU_2**	**BLEU_3**	**BLEU_4**	**ROUGE_L**	**CIDEr**	**METEOR**
HybridCNN	0.606	0.381	0.250	0.149	0.520	0.096	0.182
VGG19	0.719	0.569	0.460	0.355	0.639	0.646	0.294



Ground Truth: a man is playing a guitar
VGG19 prediction: a man is playing a guitar
HybridCNN prediction: a man is walking

Ground Truth: a man is slicing carrots
VGG19 prediction: a man is slicing a carrot
HybridCNN prediction: a woman is cutting a piece of onion

Ground Truth: three boys are synchronized dancing
VGG19 prediction: a group of people are dancing
HybridCNN prediction: a man is cutting a piece of wood

Fig. 9. A Comparison between Test-Generated Sentences and Reference Sentences.

VI. Conclusions

The automatic description of videos with natural sentences is a research problem that has been recently studied in the literature. In this study, it has been aimed to automatically obtain a natural sentence from a video. In this way, it is aimed to contribute to robotic vision tasks and help people with visual impairments. Automatic video description model should be able to express objects and events presented in the video. Automatic video description model also explains their relationships with each other in a natural sentence. To approach this problem, (Seq2Seq) model has been proposed to generate for video captioning. In this paper, the modern VGG-19 and VGG-16 CNN architectures have been used in conjunction with the LSTM. It has been aimed that the developed model has been learned to associate a variable-sized square array with a variable-sized word array. The performances of proposed models have been evaluated on the MSVD and are quantified using the BLEU, ROUGE, CIDEr and METEOR.

References

[1] World Health Organization, “Global Data on Visual Impairments 2018,” Geneve: WHO, 2018.

[2] N. Krishnamoorthy, G. Malkarnenkar, R. Mooney, K. Saenko, and S. Guadarrama, “Generating natural-language video descriptions using text-mined knowledge,” In Twenty-Seventh AAAI Conference on Artificial Intelligence, 2013.

[3] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. Forsyth, “Every picture tells a story: Generating sentences from images,” In European conference on computer vision Springer, Berlin, Heidelberg, pp. 15-29, 2010.

[4] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, “Video description generation incorporating spatio-temporal features and a soft-attention mechanism,” arXiv preprint arXiv:1502.08029, 2015.

[5] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, “Sequence to sequence-video to text,” In Proceedings of the IEEE international conference on computer vision, pp. 4534-4542, 2015.

[6] X. Li, Z. Zhou, L. Chen, and L. Gao, “Residual attention-based LSTM for video captioning,” World Wide Web, vol. 22, no. 2, pp. 621-636, 2019.

[7] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, “A dataset for movie description,” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3202-3212, 2015.

[8] N. Xu, A. Liu, Y., Wong, Y. Zhang, W. Nie, Y. Su, and M. Kankanhalli, “Dual-stream recurrent neural network for video captioning,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 8, pp. 2482-2493, 2018.

[9] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, “Dense-captioning events in videos,” In Proceedings of the IEEE international conference on computer vision, pp. 706-715, 2017.

[10] Z. Wu, X. Wang, Y. G. Jiang, H. Ye, and X. Xue, “Modeling spatial-temporal clues in a hybrid deep learning framework for video classification,” In Proceedings of the 23rd ACM international conference on Multimedia, pp. 461-470, 2015.

[11] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, “Beyond short snippets: Deep networks for video classification,” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4694-4702, 2015.

[12] X. Wang, Y. F. Wang, and W. Y. Wang, “Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning,” arXiv preprint arXiv:1804.05448, 2018.

[13] Z. Wu, T. Yao, Y. Fu, and Y. G. Jiang, “Deep learning for video classification and captioning,” In Frontiers of multimedia research, pp. 3-29, 2017.

[14] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” In Proceedings of the 40th annual meeting on association for computational linguistics, pp. 311-318, 2002.

[15] C. Y. Lin, “Rouge: A package for automatic evaluation of summaries,” Text Summarization Branches Out, 2004.

[16] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, “Cider: Consensus-based image description evaluation,” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4566-4575, 2015.

[17] P. Das, C. Xu, R. F. Doell, and J. J. Corso, “A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching,” In Proceedings of CVPR, pp. 2634–2641, 2013.

[18] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele, “Translating video content to natural language descriptions,” In IEEE International Conference on Computer Vision (ICCV), 2013.

[19] A. Torabi, C. Pal, H. Larochelle, and A. Courville, “Using descriptive video services to create a large data source for video annotation research,” arXiv preprint, 2015.

[20] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, “A dataset for movie description,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3202-3212, 2015.

[21] D. L. Chen, and W. B. Dolan, “Collecting highly parallel data for paraphrase evaluation,” In ACL: Human Language Technologies, Volume 1. Association for Computational Linguistics, pp. 190–200, 2011.

Optimized Approach in Requirements Change Management in Geographically Dispersed Environment (GDE)

Shahid N. Bhatti[1], Mohammad A. Alqarni[3]
Department of Software Engineering
College of Computer Science and Engineering
University of Jeddah, KSA

Frnaz Akbar[2]
Dept. of Software Engineering
Bahria University Islamabad, Pakistan

Amr Mohsen Jadi[4]
Dept. of Software Engineering
College of Computer Science and Engineering
University of Hail, KSA

Abdulrahman A. Alshdadi[5]
Dept. Information System and Technology
College of Computer Science and Engineering
University of Jeddah
Saudi Arabia

Abdulah J. Alzahrani[6]
Dept. of Computer Engineering
College of Computer Science and Engineering
University of Hail
KSA

***Abstract*—Managing requirements is an essential trait in engineering development process as requirements change and emerge throughout the development process. In the following research work the primary prominence is to eke out requirements that are changing frequently in geographically dispersed setup. To efficiently and effectively cope up with the changing requirements are the key to fulfill customers' requirements in geographically dispersed environment (GDE) and thus, appropriate procedural modeling is presented in this work to covenant with changing requirements to cut overall cost of the project and increase profitability by gratify the customers and the stakeholders. In the following research we have proposed an approach to tackle changing requirements in software development that are geographically dispersed and we have validated the presented procedural model through case scenario. Comprehensive systematic literature review has been performed in this section (II) to propose the efficient methodology in GDE, traits and risk and further to effectively eke out the evolving project's requirements in geographically dispersed environment. Changing requirements in geologically dispersed environment can effectively be managed if the proposed MCR model followed and it will mitigate the risk and challenges which we have to face in global software development and as well as it will cut down the overall project's cost and profitability will expectedly increase.**

Keywords—CM (Change Moderator); GDE (Geographically Dispersed Environment); MCR (Managing Changing Requirements); MCR in the GDE framework; RM (Requirements Management)

I. Introduction

Managing requirements can be described as recording, examining, organizing and concurring on requirements, and after that controlling change and conveying to applicable partners. Geographically Dispersed Environment implies the development of software system from scatters territories [3]. Managing varying requirements indicate the way of dealing the changing prerequisites with a particular true objective to satisfy the need of customers additionally it plays a vital role in the successful accomplishment of the project. Project failure risks are high if requirements are mismanaged [8]. The way toward changing prerequisites is beneficial and healthier as well as challenging at the same time.

In literature, it is accounted that due to the requirements that are frequently changing in the geographically dispersed area leads to the project failure that is very high, in these suggestions are mentioned not to continue global development at all, due to poor planning of changing requirements [5].

The present study is designed to examine how we can effectively manage the requirements that are changing frequently specifically in GDE. In the geologically dispersed setup managing changing requirements are difficult to handle due to lack of common understanding and the persistent changing in requirements. Thus it is necessary to identify the ways and appropriate methods and techniques are required to handle the changing requirements in a geographically distributed environment [2]. The following research work examines the changing requirements in the geographically distributed environment and methodology will help to tackle the change in a way to avoid failure of the project.

II. Literature Review

We have contemplated, distinctive frameworks of handling changing requirements in the Geographically Distributed Environment (GDE) from various sources. Different approaches have been projected to handle varying requirements in geologically dispersed environment. The authors in [7], conducted a review, they stated that GDE supports the advancement of plans of action and innovation in the evolution but these may lead to the project failure of the project if functional, non-functional and other requirements not considered properly, the proper understanding of change request plays an important role but the ways to tackle the risks was not highlighted. In [3], explained distinctive advantages for the advancement of global software development, the authors have used UML (Unified Modeling Language) that tackle change in the geologically distributed area, their main focus was on impact analysis of project cost how the change affects the cost.

In the research work [6] [5], the most prominent and vital advantages for the globally distributed Software are the lower advancement cost of software, the availability of professional workforces at any time in any place, proximity to the market. In addition, organizations have the opportunity for the extension of Software activities by a large number of individuals situated at various geologically distributed locales, but there are many factors that lead to failure of whole projects, but how to handle that risk not identified appropriately [10].

In the following research work [11] [14], throughout the advancement of the structure MCR their most important recognition was on verbal exchange issues during the development of software in the dispersed area, to accommodate these challenges they have suggested removing redundancy is important that will reduce the overall development time and will be beneficial for developers and as well as users or stakeholders, and the alignment of the process is significant in a globally distributed environment to fulfill the project requirements.

An international Standish group [15], examined various software projects that are 13522 in number as a consequence, “Only projects that prove to be efficacious were 29 percent, the unsuccessful percentage of projects were 18 and 53 percentage of projects were tumbled and the foremost cause is prerequisite changes that are not manage properly due to which a high failure rate is face in globally distributed software development projects”.

In context of managing changing requirements in GDE comprehensive literature review [1] [4] [9] [5] [10] [12] [13] [16] has been performed and hence identifies that 85% failure of projects are due to poor management of changing requirements in the global software development environment [7] and also identifies a novel approach to overcome these problems an improved framework has been suggested. The methodologies, the significant findings and the limitation of the used approaches have been highlighted in the Table I.

TABLE. I. DEPICTING LITERATURE REVIEW IN GDE AND MCR

Approach	Methodology	Significance	Limitations
Method applied in Management of GDE projects[3]	Proposed an ontology based Requirement.	Managing the requirements change through an architectural repository.	Requirements Change activities are not specified appropriately.
Manage Dynamic Business Process in GDE [1]	Framework based on integration of UML & CPN.	The proposed framework enhances requirements change management in business.	Utilization of Petri Net abnormal state.
A Version Control Tool for Frame-work Based Application [7]	Based on incremental technique by using GREN tool.	Bolster re-designing and system requirements change effect abridged.	Problem in GREN Wizard Source Code
Requirements Management framework in GDE [17]	Framework proposed to manage communication issues.	Covers all RE activities like change initiation, evaluation and implementation.	GDE issues and impacts not addressed appropriately.
Requirements management using XP [10]	Framework formed for RM utilizing XP in Distributed Software Development.	Manipulate the requirements in conveyed setting based on agile method extreme programming.	Cooperation isn't adequately accomplished in a disseminated setting.
Requirements Measurement Framework. [8]	To deal with MCR, metrics and indicators are used.	Based on estimation the manager of the project made choices.	Errors may occur when indicators use metrics.
Requirements tracing approach in GDE [16]	The proposed model for requirements tracing.	Stakeholders are effectively associated with each progression.	It doesn’t address the Significant issues of GDE.

III. Proposed Model for MCR in GDE

After taking into account the different issues, concerns and features in requirements management in geographically dispersed environment via comprehensive literature review (in Section II). The consensus in this is of generic essence that when the system developers no longer satisfy the clients' prerequisites (obligation) then the whole venture windup aimless for the user and the organization may confront loss and failure of project, organization in the software industry [3] [5] [7] [15].

The proposed model (Fig. 1) enhances requirements change management in the geographically dispersed environment to accommodate organizations of all kinds to manage change effectively and efficiently in the distributed environment. The framework is split into eight fundamental levels and each of these levels has particular exercises that surface throughout the procedure of managing changing requirements in the geologically distributed environment. Fig. 1, is depicting the Management of Changing Requirements in the geographically distributed environment.

A. The Request for Change in Requirements

In the projected model the MCR initiate with the demand for a modification from any associate at any circulated site, the client can also demand for change. This first stage deals with the whole statistics regarding the amendment, for example, complete illumination of requested change, the main purposes behind the alteration are, the requestor who has requested for change and so forth.

B. Impact Analysis

In 2nd phase, the requested change will send for an appraisal. It comprehends the requested amendment, for example, understanding either the modification requests for adding upgraded features to the structure or eliminating error from the system or redesign's some parts of the structure. The effect of the change on spending plan, time and other framework segments is in like manner surveyed in this stage. During the 2nd phase, request for the amendment is assessed through numerous checks.

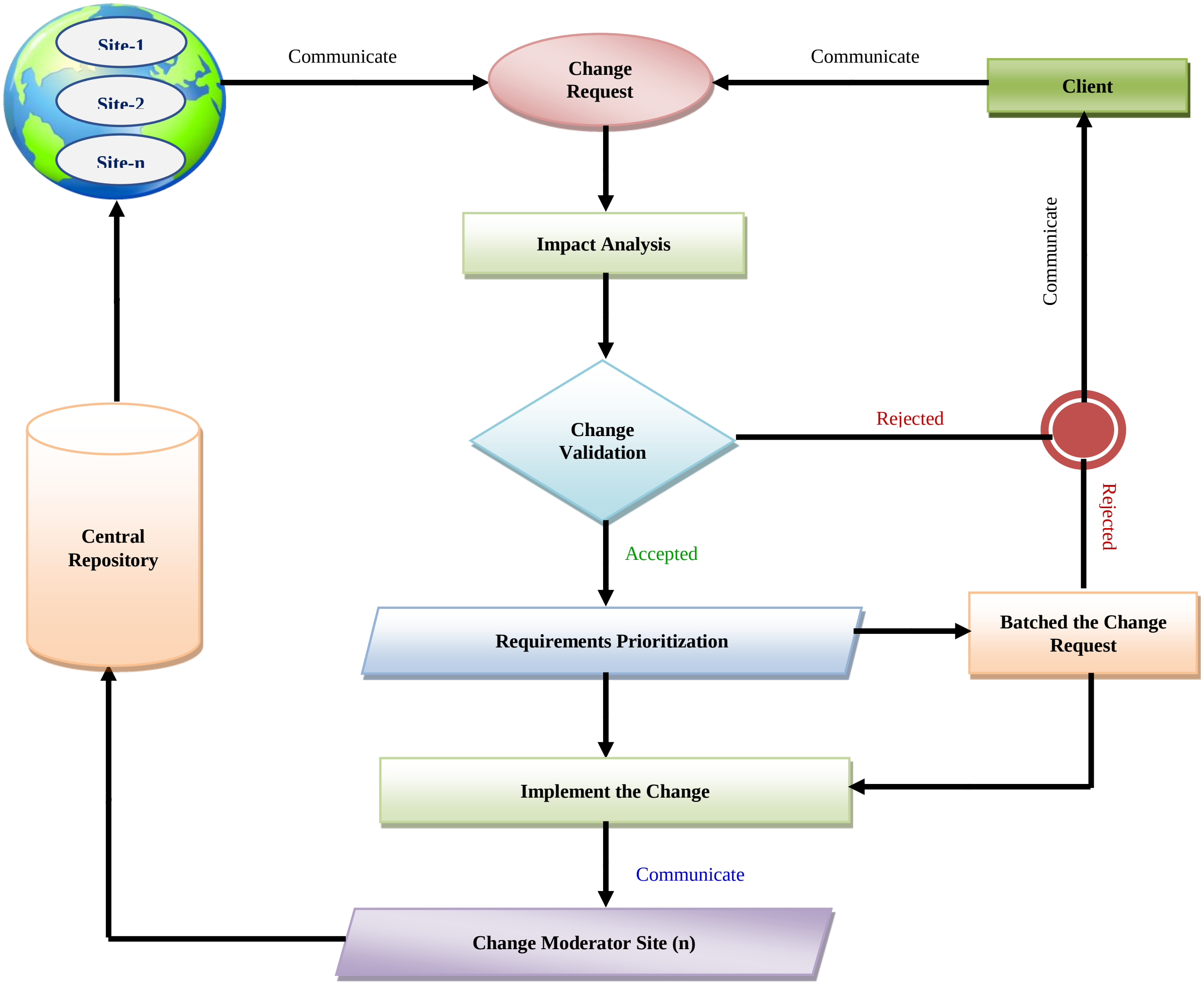


Fig. 1. Proposed Framework for Requirements Change Management in GDE.

C. Validate Changes

The reasonableness of the demanded change is assessed in the 3rd stage. The feasibility of the change is evaluated for further decision, as there are two conceivable options, whether the requested change is practical and feasible or not feasible to be executed. If the requested change isn't feasible to execute than the client informs about the impossibility of the alteration else the demand for alteration will be accepted and executed.

D. Prioritize Requirements

Organizing the requirements with the goal that the most astounding need prerequisite changes can be actualized first. Prioritization of requirements is figuring out the order of significance in the context of stakeholder as well as the developer. Prioritization of requirements may depend on individual inclination, business esteem, and the cost of usage or implementation order of the evolving prerequisite.

E. Batched the change request

A few alternate requests could be grouped together. The amendment request is assembled to execute later on according to the feasibility, because of few confinements. There are two potential outcomes with respect to clump change demands; if the requirements are easy going inside a given time, by then they will be executed else they will be rejected.

F. Change implementation

During the following stage all progressions are executed and actualize in the framework. While implementation all the Functional Requirements, Non-Functional Requirements and other Requirements considered by the implementer and by keeping the change impact in mind the implementer executes the change.

G. Inform Change Moderator

The second last step is to notify Change Moderator. The executed modification will be delivered to CM. Change Moderator will take the final decision. Change Moderator compares all previous requirements with the current change and evaluates the change.

H. Central Database

In the final phase, all the alterations are dispatched to the central database, and then all the stakeholders of the system are going to be enlightened concerning updates. Most of the system report will be revived with the change made. At long last, brand new refreshed changes will be applied to the system. Changes may reoccur at any phase at any time so the central repository of the change would be helpful for the future.

IV. Algorithm: PSEUDO

```
CRM (Change Request)
{
IF (impact analysis == accepted)
{
Prioritize Requirements;
Batch Change Request;
Implement Changes;
Change Moderator;
Update Central Repository;
}
ELSE
Inform the client & recall;
}
```

V. Validation via Case Study

We have selected xyz@ Software House Inc. (naming convention due to copy rights issue), a software program improvement corporation settled in Pakistan. The organization formed in 1990 in UK and in 2002 a branch office of the company inaugurated in Islamabad. The company develops offshore software development projects and successfully delivers a number of projects up till now. The company develops software of all types and many expert teams are available. Whenever there is a request for the change in software whether a functional change or non-functional change, the request is forward to the relevant expert team.

Let presume from the Brazil site; there is a change request in desktop app GUI; the stakeholder complaints that the GUI doesn’t meet the ease of use criteria; as they received many complaints from users that GUI is not a user friendly. So, they initiate change request the manager forward the project to the Graphic Designer team. The team assistant inquiries from the stakeholders what they actually want; in how many days they need the updated software; what were the previous flaws that they have to improve. The teams setup the meeting and decide what to change and assign the tasks to the workforces. Each team member performs their assigned tasks and informs the managerial head and the improved software forward to the stakeholders, but they face many issues during this change. As sometimes one single change affects many components and the overall cost, time for the development also increase. Sometimes it happens that the change request cost increased than the actual cost of software; another technical and non-technical risk also arises during the change; like difficult to synchronize, the stakeholder or the user dissatisfies with the updated change.

If the proposed framework is followed by the company, all above mentioned issues can easily be tackled. As if there is a change request first of all a meeting should be conduct to discuss the impact analysis of the requested change on the whole project, then validate that change if the change is feasible or not, if not then inform the requester. If the requested change is acceptable, then the project further handover to the relevant team, the team members have to conduct the meeting for prioritizing the requirements, and batch the change request that cannot be implemented right now due to some limitation and then tasks will be assigned to experts they will implement the change and inform the change moderator, the change moderator will review the change and update the central repository. Finally change requester will be informed about the change and reviews by the change requester should be kept safe.

The proposed model efficiently manages the changing requirements in the GDE environment and avoids the project failure percentage. Fig. 2 shows the graph of successful projects by applying the proposed model (Fig. 1), and Table II shows how effectively requirements change managed by following the proposed framework.

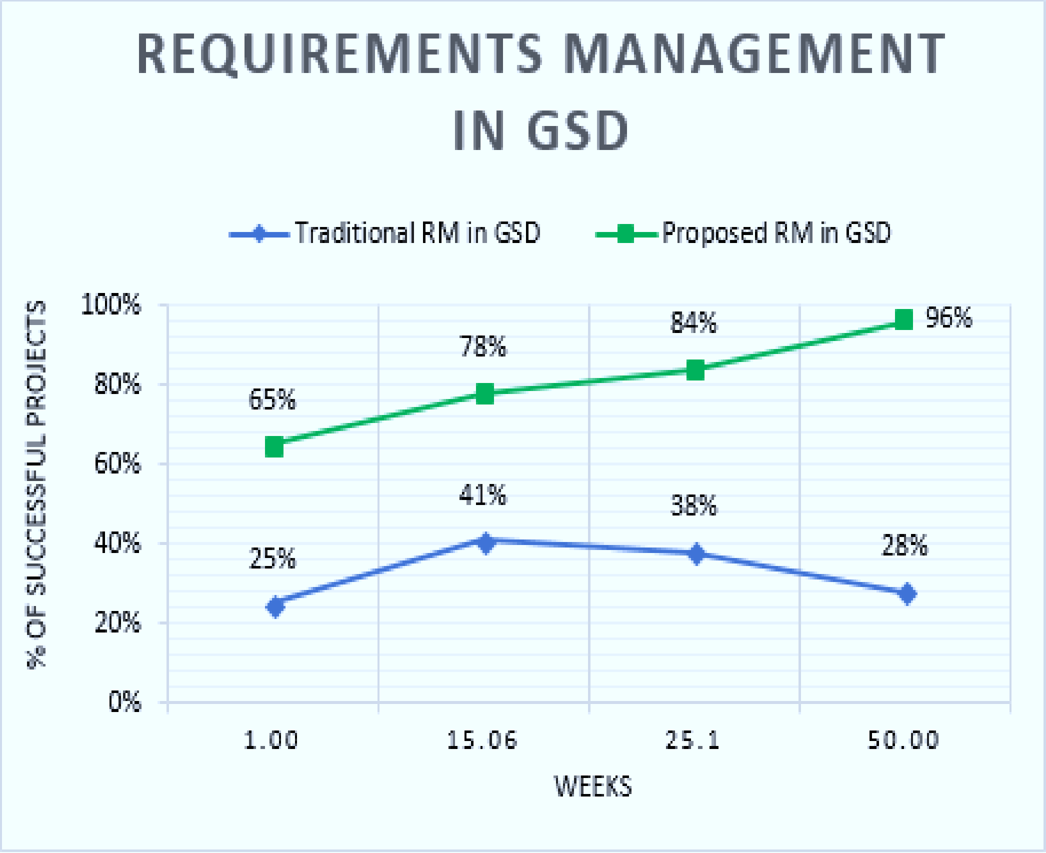


Fig. 2. The Percentage of Projects that Met Objectives.

TABLE. II. MCR IN GDE TRADITIONAL VS. PROPOSED MODEL

Requirements Management in GDE % of Successful Projects		
Change Management	**Traditional RM in GDE**	**Proposed RM in GDE**
Poor	25%	0%
Fair	70%	1%
Good	5%	7%
Excellent	0%	92%

VI. Conclusion

The requirements emerge, change throughout the software development process and requirements are needed to be prioritized and hence managed with utmost priority, especially when the scenario is that of Global Software Development (GSD) and especially geographically distributed environment (GDE). The essential intention of the following research work is to formalize a framework that efficiently oversees managing change requirements (MCR) in the GDE environment. As efficiently handling MCR in the GDE environment save cost, time and as well as ensure the availability of resources, highlighted in detail in literature review of this paper. A complete set of steps are presented, each step if carefully process the failure risk can be lessened. The proposed framework will lead to efficiently and effectively manage the change in requirements in the GDE environment.

Although it is suggested here that further work is required to handle communication risks and quality maintenance of requirements in the GDE environment for more effective results. Communication issues in the GDE environment also result in project failure so for better understanding of change there should be check and balance. Quality is the first priority of every stakeholder, so by centering quality maintenance in the GDE environment, the project reliability can be increased.

References

[1] Minhas, Nasir., & Qurat.,& Zafar.,& Atika, Zulfiqar. (2014).“An Improved Framework for Requirement Change Management in Global Software Development”. Journal of Software Engineering and Applications.7, 779-790.

[2] Bhatti Shahid., & Usman Maria., & Jadi Amir. (2015). “Validation to the Requirement Elicitation Framework via Metrics”. ACM SIGSOFT, USA, 40, 1-7.

[3] Hussain, Waqar. (2010). “Requirements Change Management in Global Software Development: A Case Study in Pakistan.”Retrieved from https://pdfs.semanticscholar.org/41cb/6b1d7b2b03f1214282c812c74ad1e65860d6.pdf

[4] Aneesa R. Asghar, Shahid N. Bhatti, (2017). “The Impact of Analytical Assessment of Requirements Prioritization Models: An Empirical Study” International Journal of Advanced Computer Science and Applications (IJACSA), 8(2), 2017.

[5] Holmstrom, Helenna., & Conchr, Eoen., &Agerfalkh, Paar., & Fitzgerald, Brian. (2006). “Global Software Development Challenges: A Case Study on Temporal, Geographical and Socio-Cultural Distance”, ICGSE, Brazil, 2006. Brazil: Costão do Santinho, Florianópolis.

[6] Rida, Shahid Bhatti. (2017). “Impact and Challenges of Requirements Elicitation & Prioritization in Quality to Agile Process: Scrum as a Case Scenario, Conference: IEEE, International Conference on Communication Technologies (ComTech-2017), Rawalpindi, Pakistan.

[7] Mighetti, Juan., & Hadad, Graciela. (2016). “A Requirements Engineering Process Adapted to Global Software Development, Argentina”, 2016. Argentina: CLEI Electronic Journal.

[8] Sehrish Alam, Amr Mohsen Jadi. (2017). “Impact and Challenges of Requirement Engineering in Agile Methodologies: A Systematic Review” International Journal of Advanced Computer Science and Applications(IJACSA), 8(4), 2017.

[9] Jayatilleke, Shalinka., & Lai, Richard. (2017). “A systematic review of requirements change management”, SIIA, Bundoora, 2017. Australia: INFSOF.

[10] Aneesa R., Atika. (2016). “Role of Requirements Elicitation & Prioritization to Optimize Quality in Scrum Agile Development” International Journal of Advanced Computer Science and Applications (ijacsa), 7(12), 2016.

[11] Ahmad, Zahoor., & Hussain, Mussarat., & Rehman., & Qamar, Usman., & Afzal, Muhammad. (2015). “Impact minimization of requirements change in software project through requirements classification, International Conference on Ubiquitous Information Management and Communication”, Bali, 2015. Indonesia: ACM.

[12] Amna Sadiq, Makkia Abbasi. (2017). “Requirements Prioritization, management Techniques: An Empirical Study”. International Conference on Computing and Mathematical Sciences (ICCMS 2017), Invent, Innovate and Integrate for Socioeconomic Development. 25th and 26th February, 2017, ICCMS 2017.

[13] Raffo, David., & Wakeland, Wayne., & Setamint, Siri. (2006). “Planning and improving global software development process using simulation. Proceeding”, 18, 8-14.

[14] Sabahat, Nosheen., &Iqbal,Faiza., &Azam, Farouqee.,& Javed, Younas. (2010). “An Iterative Approach for Global Requirements Elicitation: A Case Study Analysis”, 2010.“International Conference on Electronics and Information Engineering (ICEIE)”.

[15] Prikladnicki, Rafaeel., & Audy, Jorge., & Evaristo, Roberto. (2014). “Global Software Development in Practice Lessons Learned”,8,267-281.

[16] Ali, Shahid, Tayyab, “Impact of Story Point Estimation on Product using Metrics in Scrum Development Process” International Journal of Advanced Computer Science and Applications(IJACSA), 8(4), 2017.

[17] Khan, Arif., &Basri, Shuib. "A propose framework for requirement Changes Management in Global Software Development", 2012.“International Conference on Computer & Information Science (ICCIS).

Data Mining for Student Advising

Hosam Alhakami[1], Tahani Alsubait[2], Abdullah Aljarallah[3]
College of Computer and Information Systems
Umm Al-Qura University, Makkah, Saudi Arabia

***Abstract*—This paper illustrates how to use data mining techniques to help in advising students and predicting their academic performance. Data mining is used to get previously unknown, hidden and perhaps vital knowledge from a large amount of data. It combines domain knowledge, advanced analytical skills, and a vast knowledge base to reveal hidden patterns and trends that are applicable in virtually any sector ranging from engineering to medicine, to business. However, it is possible for educational institutes to use data mining to find useful information from their databases. This is usually called Educational Data Mining (EDM). Advancing the field of EDM with new data analysis techniques and new machine learning algorithms is vital. Classification and clustering techniques will be used in this project to study and analyse student performance. The key importance of this project is that it discusses different data mining techniques in the literature review to study student behaviour depending upon their performance. We tried to identify the most suitable algorithms from the existing research methods to predict the success of students. Various data mining approaches were discussed and their results were evaluated. In this paper, the J48 algorithm was applied to the data set, gathered from Umm Al-Qura University in Makkah.**

Keywords—Data mining; performance prediction; student analytics; academic advising; classification algorithms; decision tree; J48; neural network; Weka

I. Introduction

Educational institutions are very important for the socio-economic development of a country and students are the building blocks of any educational institute. They are very important for society because they are responsible for the future development. Depending upon the role of students in a country, their performance measure is an important aspect for any institution. The performance of a student is not only important to the institution, but it also affects the corporate sectors and job markets. High ranked universities produce great leaders in their particular domains because they put their resources to judge student's abilities and predict their field depending upon their performance. Several factors influence how students perform academically, and they include socio-economic factors as well as other environmental variables [1]. The impact of such factors can be better managed when people know about them and how they affect the performance of students. The tools, approaches, as well as the investigation that is designed to extract meaning from huge data sources of data that people learning activities generate automatically in an educational environment, is referred to as Educational Data Mining [2]. Investigations on educational mining have been focused on, to a large extent, in recent times. There is a special emphasis on the area of explaining and predicting academic performance. In fact, the big data contained in educational databases makes it more difficult to predict the performance of students. It is suggested that the model based on institutional internal databases and external open data sources performs better than the model based on only institutional internal databases [3]. Nevertheless, it is crucial for one to be able to predict the performance of students in an educational environment. Every academic institution has a long-term goal of ensuring that the success of students increases. The ability of an educational institution to predict the academic performance of students on time before their final examination makes it possible to put in additional efforts to make arrangements to help students that are not performing well and ensure they succeed. However, it is possible to aid improvement in courses by identifying the characteristics that influence the success rate of students in these courses. Investigators have the rare privilege of studying students' learning behaviours, as well as the methods that can help achieve success by using technology-based educational tools that are developed recently and by applying quality standards.

This paper aims to study the performance of students using different data mining techniques such as classification and clustering and provides a suitable technique that could be used by student advisors. This study will help the universities to improve the performance of the students. It introduces student marks prediction models using predictive model approaches based on student behaviour. We used data sources from Umm Al-Qura University that were used in practical settings to predict students' academic performance. Also, it focuses on identifying the variables that can be used to predict the future performance of students.

II. Background

The explosion of database management systems has led to a massive collection of every kind of information nowadays [4]. To date, the available information from military intelligence, text reports, satellite pictures, scientific data, and business transactions, is more than can be handled. It is no longer enough to retrieve information for decision-making [5]. There is currently the creation of new needs that will assist in making better managerial decisions. The needs include the discovery of patterns in raw data, extracting the significance of stored information and automatic summarization of data.

A. Data Mining Techniques

Data mining is also commonly referred to as Knowledge Discovery in Databases (KDD), referring to the knowledge discovery process, knowledge extraction, knowledge mining from data or data/pattern analysis [5]. It refers to the nontrivial extraction of implicit, previously unknown and possibly valuable information from data in databases [6]. Ramageri et al. [7] explain that data mining can be described as a process through which useful information and patterns are extracted from big data.

According to Ramaraj et al. [8], classification refers to the task of generalising a known structure and applying it to a new one. It is possible for data mining classification techniques to process big data [9]. It helps to predict categorical class labels and categorises data by the training set and class labels, and it is also useful for categorising data that is available recently. Nikam [10] stated that the classification procedure is an established technique, which constantly makes those types of decisions in new situations. Decision Tree, Neural Networks, Naïve Bayesian Classification, Support Vector Machines, and K-Nearest Neighbour are common algorithms used to classify data.

1) A decision tree classifier: It is a classifier that uses the instance space's recursive partition. It is made of nodes that form a rooted tree, which means that it is a directed tree with a node referred to as roots, which have no incoming edges [8]. Intermediate nodes generate outgoing edges, and they test the nodes after performing gates. It consists of a decision tree that is generated according to instances. Nodes that do not consist of an outgoing branch are referred to as terminal or decision nodes [11]. Individual internal nodes split the instance space into two or more sub-spaces in a decision tree. The internal nodes, as well as the root, relate to features, while leaf nodes relate to classes [8]. All in all, there is an outgoing branch of individual non-leaf nodes, for every probable value of the characteristics that are related to the node. Sequential nodes are visited pending the time that a leaf node is reached, when using a decision tree to decide the class for a new instance, beginning with the root [8]. A test is applied at the root node and every internal node. The crisscrossed branch, as well as the next node visited, is determined by the result of the test.

2) Neural networks: It uses the gradient descent technique of the biological nervous system that has several interrelated processing elements. Such elements are referred to as neurons. The learned network's operability is improved using the rules that are extracted from the trained neural network [11]. Put differently; neural networks can be described as an emulation of the biological neural system [12]. It comprises an interconnected group of artificial neurons as well as processes information through a connectionist technique to computation. Generally, neural networks are adaptive systems in which internal or external information flowing through the network during the learning phase changes its structure [12].

3) Naïve bayesian classification: A Naïve Bayes classier refers to a simple probabilistic classier that functions with Bayes theorem (from Bayesian statistics) that has a strong (naïve) independence assumption.

Put differently; a naïve Bayes classifier assumes the presence or absence of a specific attribute of a class is not related to any other attribute's presence or absence [13]. It does not matter if these attributes rely on one another or the presence of other attributes. It is possible to train naïve Bayes classifiers efficiently in a supervised learning system, based on the probability model's exact nature. It is the technique of maximum probability that naïve Bayes models use an estimation parameter in several practical applications. This type of classifier has worked well in several real-world complex situations according to Irina et al. [14], even though it has naïve designs as well as over-simplified presumptions.

4) Support Vector Machines (SVM): The first person to introduce the modern method of classification called the support vector machine is Vapnik [15]. It is commonly used in bioinformatics because it is highly accurate, and it can handle high-dimensional data such as flexibility in modelling different data sources, as well as gene expression [16]. It belongs to the general group of kernel techniques [17]. The support vector machine (SVM) has been an effective technique for general pattern recognition, classification as well as regression. It can be said to be an excellent classifier as it has a high generalisation performance that does not require the addition of previous knowledge, regardless of whether it has an extremely high dimension of the input space. The SVM aims at looking for the best classification function that helps to differentiate between members of both classes in the training data [8]. It is possible to use geometrics to determine the best classification function. Regarding a dataset that can be separated linearly, a linear classification function matches a separating hyperplane f(x) which goes through the middle of both classes, separating them [11].

5) K-Nearest neighbour classifiers: This type of classifier is based on learning by analogy. It is the n-dimensional numeric characteristic that describes the training samples. Every sample signifies a point in n-dimensional space [8]. Along these lines, the entire training samples are stored in n-dimensional pattern space. The k-nearest neighbour classifier looks for the k training samples' training samples which are closest to the unknown sample when it is given an unknown sample. It is the Euclidean distance that determines the closeness. Nearest neighbour classifiers give every attribute equal weight.

Also, it is possible to use it for prediction; in other words, it can be used to return a real-valued prediction for a given sample that is unknown. The algorithm of k-nearest neighbours is among the easiest machine learning algorithms. It is when there is a majority vote of the neighbours of an object [8]. It is crucial to choose an appropriate I value when using a k-nearest neighbour algorithm [8].

B. *Analytical Tools Weka*

The formatting of datasets in Weka should be in the ARFF or CSV format. These would be used automatically in Weka Explorer when a particular file is not recognised. Data can be imported from a database using the facilities contained in the pre-process panel; this data utilises a filtering algorithm when it comes to pre-processing. It is possible to convert this data using the filters, which allows the deletion of cases and assigns based on specified principles.

WEKA tool's Classification algorithm is used to carry out experiments on the data set. The first step involves searching for an aggregate number of cases of the particular data using the j48 classification algorithm as well as Naïve Bayes. The following step requires the experiment to conduct the cost analysis as well as find out the correctness of the Classification.

III. Related Work

Johnson (2018) [18] stated that there is a high demand for enrolments in computer science, and the graduations of successful computer science undergraduates are, without a doubt, very significant. He, therefore, proposed building upon the current data mining and modelling, learning analytics, and machine learning applications for predicting the success of students that goes beyond retention. To Khare et al. (2018) [19], educational data mining (EDM) refers to an applied field of research, which combines data mining, statistics, and machine learning in the field of education, but not restricted to MOOCs, intelligent tutoring systems, universities and schools. They decided to explore the essence of data mining in the online education setting and discover its ways of improving the learning experience of the student. They intend to achieve their objective by reviewing some of the basic data mining algorithms used in education and the innovations that will come up in the future. The proposal presented by Brooks (2013) [20] states that EDM community research's distinction emanates from intelligent tutoring systems, which is the interaction between the student, domain material and the system, whereas, the focus of the learning analytics researchers is on enterprise learning system such as classroom management systems, which pile up data from all the courses.

However, according to Thomas and Gelan [21], the definition of learning analytics is that they collect, measure, evaluate and report data regarding the learners in their context to understand and optimise learning and learning environments. They believe that learning analytics presents some potential opportunities because, through it, teachers can obtain valuable data on learners that are succeeding and failing. On the other hand [22] stated that students that take transfer in community colleges face challenges in their pursuit of bachelor's degrees, and this usually leads to credit loss. To them, this may consequently reduce the students' chances of completing their credential, and increase the costs and time for the students, their families, and taxpayers.

According to Lacefield [20], accountability appears to be permanently rooted in the environment of K-12, like the expectation of delivering quality education to children of school and adolescents. However, this expectation has failed repeatedly and has drawn the attention of policymakers and the public to the drawbacks of major accountability systems. Therefore, they tried to show how to use the predictive analytics applied to school student system (SIS) records, to make advising of students and activities such as mentoring or coaching at-risk students easier.

The argument of Dash and Vaidhehi [23] stated that academic advising requires much time, responsibility and skills. They believe that it is necessary for the computerised advising system to be forthcoming so that human advisors can be assisted effectively. Additionally, Olaniyi [24] said that Educational Data Mining (EDM) focuses on developing and modelling techniques that discover knowledge from data emanating from educational environments. Moreover, based on the perception of Pal, and Chaurasia [25], the consumption of alcohol in higher education institutions is not new; India's legal drinking age is 18 years, but it is dangerous for underage students and those that are 18 years and above to drink heavily. Therefore, four popular data mining algorithms; REP Tree, Bagging, Sequential Minimal Optimisation (SMO) and Decision Table (DT), obtained from a rule-based classifier or a decision tree to improve academic performance's efficiency in the educational institutions for alcohol-consuming students were discussed.

El-Halees and Abu-Zaid [26] stated that presently, websites are perceived as a vital tool in several real-life applications including, entertainment, industry, education, and business, and this has brought many concerns regarding the quality of these websites. Nevertheless, Hussain [27] pointed out that Google has one million search queries every one minute on the internet; more than two million emails are sent, there are 100,000 tweets, thousands of photos are uploaded, and much more traffic. To them, terabytes of data, which has a grade value that can shape higher education institutions generate the future of nations. Moreover, Yadav et al. (2012) [27] believe that Knowledge Discovery and Data Mining (KDD) is a multifaceted discipline that focuses on the methods used to extract valuable knowledge from data. Education's quality can be increased using this knowledge.

According to Tair and El-Halees [28], educational mining focuses on developing methods of knowledge discovery using data collected from the field of education. To Baepler and Murdoch [29], the areas of academic analytics and educational data mining have experienced rapid development, and the outcome has resulted in new potentials for collecting, analysing and presenting student data.

Weakley et al. [30] said there are divergent views that the use of predictive measures violates a professional ethical principle to develop a comprehensive understanding of their advice, according to professional advisers at the Public Higher Education Institute, which may contribute to the enrichment of content in the exploration of Educational data mining from a social and ethical perspective.

IV. Methodology and Research Design

A. Procedure

This project involves some major steps, including:

- Step 1: Collect student-related information from the university for at least five years for the complete information regarding the dataset used.
- Step 2: Clean and verify the student information collected from Step 1. Microsoft Excel functionalities such as data filtering, data sorting, and so on, would be used to verify, validate and clean the data manually.
- Step 3: To allow us to study the student's performance, once the data is cleaned, verified and categorised. Step

1 to Step 3 will be repeated for the entire datasets received from the university.

- Step 4: To use the Weka data mining tool to perform the classifications, to analyse the appropriate algorithm.
- Step 5: Analyse the performance of the algorithms from the results received from step 4.
- Step 6: To evaluate the recommended classification method in this project as well as the prediction results with experienced lecturers. This method is key to ensure that the suggestions we provided are appropriate for real-life usage.

B. *Classification Algorithms*

The following classification algorithms were compared according to their performance on the dataset:

Decision tree (J48): It is a predictive machine-learning model whose function is to determine the new sample’s target value using various attribute values of the available data. The decision tree’s internal nodes represent the various attributes. The branches between the nodes indicate these attributes’ possible values in the experimental samples, and the final value of the dependent variable is indicated by the terminal nodes [31]. J48 algorithm has been used to generate decision tree using the Weka tool.

Naive Bayes: This type of classifier is based on the rule of Bayes that expresses an event’s possibility before there is a clear proof, as well as an event’s possibility after the proof becomes apparent [32]. There are several reasons for using this classifier, and they include building the easiest way without requiring any complicated iterative parameter evaluation schemes.

C. *Performance Measurement Factors*

A brief discussion of the performance measurement factors is provided below:

TP Ratio: TP stands for True positive. The number of dataset rows that are predicted as positive that are truly positive is known as TP ratio.

FP Ratio: FP stands for False positive. The number of dataset rows that are predicted as positive that are truly negative is known as FP ratio.

Accuracy: This is measured based on the ratio of the correct observation over the particular dataset’s overall observations.

Precession: This is measured based on the ratio of the positive observation that has been predicted accurately over the overall positive predicted observations.

Recall: It is measured based on the ratio of the positive observation that has been predicted over all the observations in actual yes class.

F-measure: This refers to an average between the recall and accuracy.

Classification Matrix: This refers to a table whose primary purpose is to represent the performance of a particular classification model that can be any algorithm.

D. *Dataset*

The dataset was gathered from the information of graduate students from Umm Al-Qura University in the last 5 years in Makkah. 26711 student records were used for training and 11960 students for testing. The total number of students is 38671. Some records with incomplete data were discarded, the total number of gathered records is 59699.

The data fields in the dataset are provided below:

- NATIONALITY: This is the field column in the dataset which provides the student nationality is where they are a legal citizen.
- SCHOOL GOVERNATE: This is the field column in the dataset which provides the student school governate is an administrative division of a country.
- GRADE: This is the field column in the dataset which provides the student Final student grade at the university (Excellent, Very Good, Good, Pass).
- TAHSILI MARK: This is the field column in the dataset which provides the student mark in Tahsili exam (i.e., standardised national exam for student’s academic performance).
- QDRAT_MARK: This is the field column in the dataset which provides the student mark in Qdrat exam (i.e., standardised national exam for student’s skills measurement).
- CAMPUS: This is the field column in the dataset which provides the grounds and buildings of a university, college.
- GENDER: This is the field column in the dataset which provides the student state of being male or female.
- SCHOOL AVERAGE: This is the field column in the dataset which provides the student school average.
- SCHOOL BRANCH: This is the field column in the dataset which provides the student internal specialization in school. Many of these specializations are special to Makkah district such as Dar Al Tawheed, Literary Section, Memorization of the Koran, scientific department, Dar Al Hadith, scientific department Courses, Commercial Secondary, Literary Section, Al - Haram Institute, Literary Section Courses, Noor Institutes, Holy Quran Institute for National Guard, Health Institute, Visual impairment, Teachers Training Institute, Secondary Teacher Institutes, Secondary professional, Secondary Alsolatyh, Scientific Institute.
- FACULTY: This is the field column in the dataset which provides the student colleges at universities.
- AGE: This is the field column in the dataset which provides the student age.

V. Results and Discussion

A. Weka Analysis

The Weka result for decision tree (J48) algorithm is shown in Fig. 1.

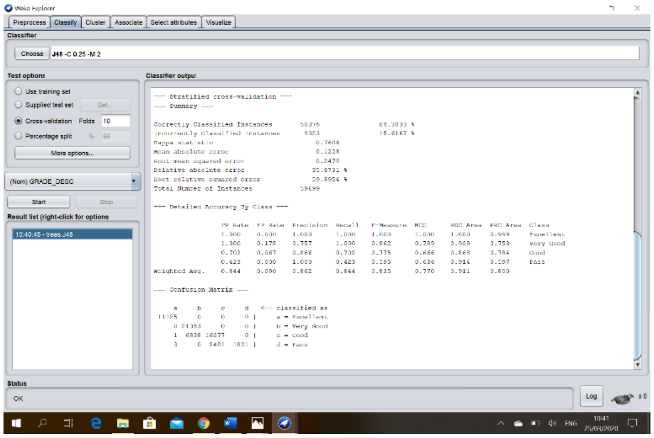
As shown in Fig. 1 for J48, the percentage of correctly classified instances is 84.38%, while the percentage of the incorrectly classified instances is 15.61%. On the other hand, as shown in Fig. 2 for Naive Bayes algorithm, the percentage of correctly and incorrectly classified instances are 46.68% and 53.31%, respectively. In this scientific paper, the logarithm of J48 is chosen compared to the logarithm of Naive Bayes, due to the fact that it is correctly classified as the highest in the algorithm of J48.

B. Analysis of the Factors that affect Students' Performance

The basic dataset analysis was performed using pivot table functionality and the basic statistical functions of the MS Excel.

Fig. 3 shows that the percentage of female is the highest in the excellent score (78%) and the score is very good for 65% females, while male students are 22% excellent and 35% very good.

Table I contains four elements (PASS, GOOD, VERY GOOD and EXCELLENT) to describe the student's performance in the final rate (GPA) based on gender.



J48 -C 0.25 -M 2

Folds 10

(Nom) GRADE_DESC

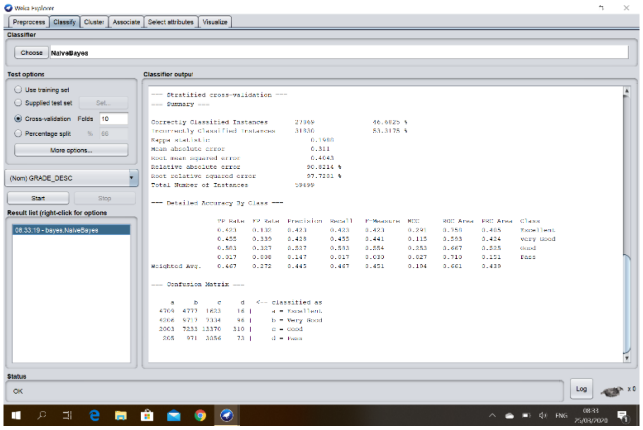
10:40:45 - trees.J48

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       50376               84.3833 %
Incorrectly Classified Instances      9323               15.6167 %
Kappa statistic                          0.7606
Mean absolute error                      0.1228
Root mean squared error                  0.2478
Relative absolute error                 35.8731 %
Root relative squared error             59.8954 %
Total Number of Instances            59699

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     0.999     Excellent
                 1.000    0.178    0.757      1.000    0.862      0.789    0.909     0.753     Very Good
                 0.702    0.067    0.866      0.702    0.775      0.666    0.868     0.784     Good
                 0.423    0.000    1.000      0.423    0.595      0.636    0.914     0.597     Pass
Weighted Avg.    0.844    0.090    0.862      0.844    0.835      0.770    0.911     0.800

=== Confusion Matrix ===

     a     b     c     d   <-- classified as
 11125     0     0     0 |     a = Excellent
     0 21353     0     0 |     b = Very Good
     1  6838 16077     0 |     c = Good
     3     0  2481  1821 |     d = Pass
```

OK

10:41
25/03/2020

Fig. 1. Use Training Set to Learn the Algorithm J48.



NaiveBayes

Folds 10

(Nom) GRADE_DESC

08:33:19 - bayes.NaiveBayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       27869               46.6825 %
Incorrectly Classified Instances     31830               53.3175 %
Kappa statistic                          0.1988
Mean absolute error                      0.311
Root mean squared error                  0.4043
Relative absolute error                 90.8214 %
Root relative squared error             97.7201 %
Total Number of Instances            59699

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.423    0.132    0.423      0.423    0.423      0.291    0.758     0.405     Excellent
                 0.455    0.339    0.428      0.455    0.441      0.115    0.593     0.424     Very Good
                 0.583    0.327    0.527      0.583    0.554      0.253    0.667     0.525     Good
                 0.017    0.008    0.147      0.017    0.030      0.027    0.710     0.151     Pass
Weighted Avg.    0.467    0.272    0.445      0.467    0.451      0.194    0.661     0.439

=== Confusion Matrix ===

     a     b     c     d   <-- classified as
  4709  4777  1623    16 |     a = Excellent
  4206  9717  7334    96 |     b = Very Good
  2003  7233 13370   310 |     c = Good
   205   971  3056    73 |     d = Pass
```

OK

08:33
25/03/2020

Fig. 2. Use Training Set to Learn the Algorithm Naïve Bayes.

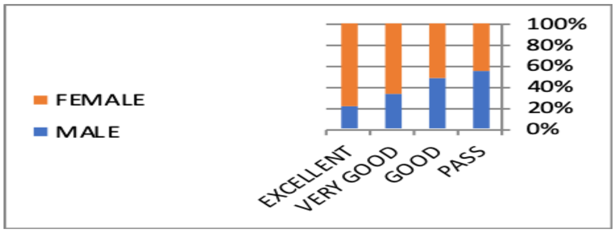


Fig. 3. The Grade for all Student based on Gender.

TABLE I. The Grade for All Student based on Gender

GENDER_DESC	***EXCELLENT***	***VERY GOOD***	***GOOD***	***PASS***
MALE	2458	7245	11085	2341
FEMALE	8667	14108	11831	1964

Depending on the percentage, we can focus on the development and training process on male students and motivate them through the academic supervisor.

Fig. 4 indicates that the percentage of Jamoum and Laith students is the highest in the good grade of (39%). The percentage of Business and IT is the highest in the very good grade of 40%. The percentage of Islamic is the highest in the good grade of 45%. The percentage of Education is the highest in the very good grade of 48.5%. The percentage in pass grade is 1%. The percentage of Medical is the highest in the very good grade of (49%)

Students studying in Laith and Jaumum colleges are the lowest in the percentage of the university average, since they have studied high school in Makkah, suggesting that we should focus on the development and training process of students and motivate them through academic supervision compared to other colleges.

Fig. 5 indicates that the percentage of Memorization of the Koran student is the highest in the excellent grade of 46%. The percentage of the Literary Section is the highest in the good grade of 44%. The percentage of the scientific department is the highest in the very good grade of 37%. The percentage of scientific department Courses is the highest in a very good grade of 42%. The percentage of Literary Section Courses is the highest in the very good grade of 41%. The percentage of Scientific Institute is the highest in the good grade of 42%. The percentage of other departments is the highest in the excellent grade of 32%.

The percentage shows that we can focus on the development and training process on the student who studied in high school in the literary section or scientific department and motivate them through the academic supervisor.

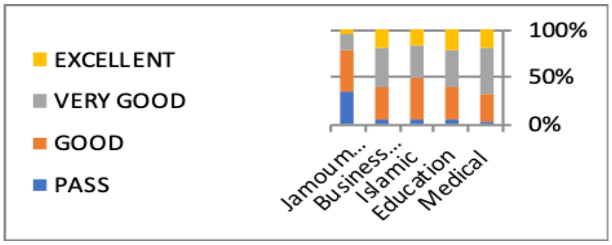


Fig. 4. The Grade based on College for Students who Study High School in Makkah.

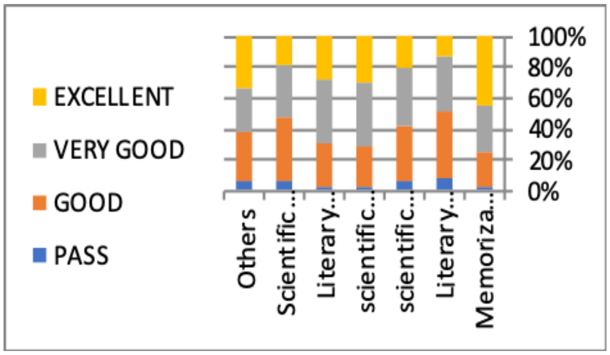


Fig. 5. The grade for student based on department in the high school who study high school in Makkah.

Fig. 6 indicates that the percentage of the Tahsili exam (100-95%) is the highest in the excellent grade of 88% and the grade of very good is 12%. The percentage of the Tahsili exam (94-90%) is the highest in the excellent grade of 60%. The percentage of the Tahsili exam (89-80%) is the highest in the excellent grade of 46% and the grade of very good is 34%. The percentage of the Tahsili exam (79-70%) is the highest in the very good grade of 42% and the grade of good is 28%.

Students who have a rate of (38-54%) and a rate of (69-55%) in the Tahsili exam are the lowest in the percentage in the university average, suggesting that we should focus on the process of developing and training students and motivate them through the academic supervisor, at the beginning of studying at the university.

Fig. 7 indicates that the percentage of the Qdrat exam (100-95%) is the highest in the excellent grade of 69%. The percentage of the Qdrat exam (94-90%) is the highest in the excellent grade of 55%. The percentage of the Qdrat exam (89-80%) is the highest in the excellent grade of 37%. The percentage of the Qdrat exam (79-70%) is the highest in the very good grade of 42%.

Understudies who have a percentage of (42-54%) and (69-55%) in the Qdrat test are the least in the rate in the GPA, we can help them during the time spent creating and preparing understudies and spur them through the academic supervisor, toward the start of learning at the college.

Fig. 8 indicates that the percentage of student age (16-22 y) is the highest in the excellent grade of 13% and the grade of very good is 11%. The percentage of student age (23-29 y) is the highest in the excellent grade of 78%. The percentage of student age (30-39 y) is the highest in the passing grade of 21%. The percentage of student age (40-54 y) is the highest in a very good grade of 3%.

The largest percentage of students is confined to the age group (23-29 y), and we can focus on students with the age group (30-39 y) as the lower percentage increases in it, and this helps the academic supervisor to guide students and increase training and development for them.

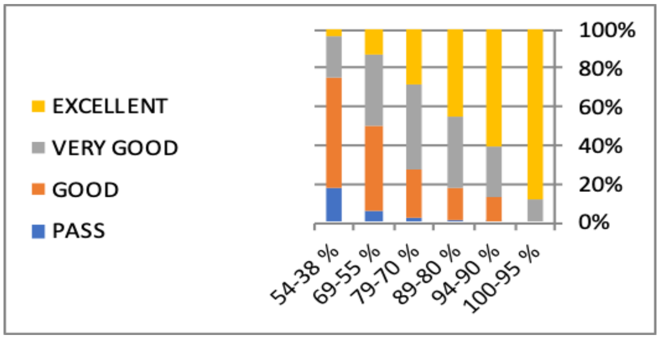


Fig. 6. The grade for all student who take the Tahsili exam.

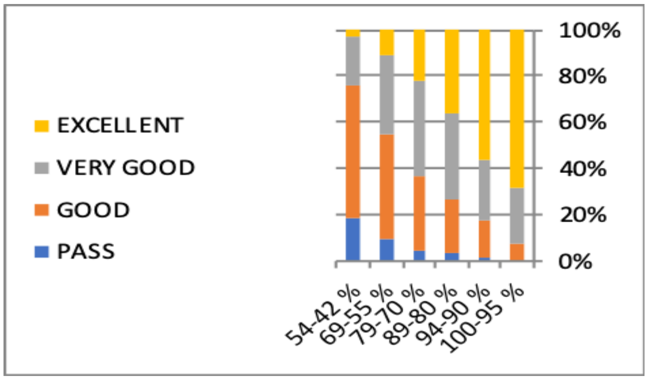


Fig. 7. The grade for all student who take the Qdrat exam.

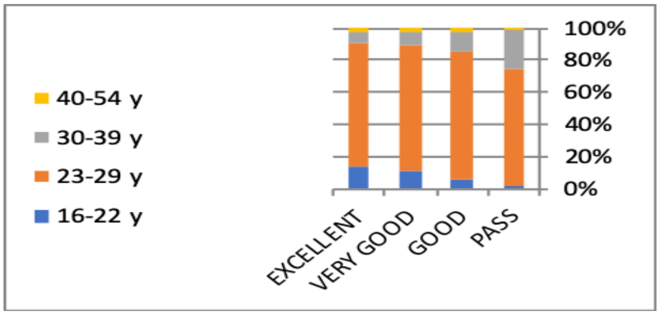


Fig. 8. The grade for all student based on age.

VI. Conclusion and Future Work

In conclusion, the decision tree (J48) classification algorithm was used to analyse student performance. Moreover, the project introduced a model that can predict the college and the GPA of a specific student by analysing the exams and other features. The accuracy of the prediction is high because it has identified that GPA and extra tests are not the only factors that affect the final results of the student. This project has identified additional factors that can be influencing student performance which are School, Sex, Age, Nationality, and City. Additional factors that this project has identified as factors influencing the student performance, as would be expected, are High School Ratio, Qdrat exam, Tahsili Exam, and department in high school.

References

[1] Surjeet Kumar Yadav, Brijesh Bharadwaj, and Saurabh Pal. “Data mining applications: A comparative study for predicting student’s performance”. In: arXiv preprint arXiv:1202.4815 (2012).

[2] P Nithya, B Umamaheswari, and A Umadevi. “A survey on educational data mining in field of education”. In: International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 5.1 (2016), pp. 69–78.

[3] Farhana Sarker, Thanassis Tiropanis, and Hugh C Davis. “Students’ performance prediction by using institutional internal and external open data sources”. Accessed from: http://eprints.soton.ac.uk/353532/1 /Students'%20mark%20prediction%20model.pdf (2013).

[4] Aliyar, F. Database Needs and Data Mining. International Research Journal of Management Science and Technology.2010.

[5] Atul Goyal, Dinesh Kumar, and Ms Sunita Baniwal. “Performance Analysis Of K-Mean Algorithm Using Markov Chain Lloyd”. Internatinal Journal of Research Review in Engineering Science & Technology, 4 (1) (2015), pp. 7–11.

[6] Osmar R Za¨ıane. “Principles of knowledge discovery in databases”. In: Department of Computing Science, University of Alberta 20 (1999).

[7] Ramageri, B. Data Mining Techniques And Applications . Indian Journal of Computer Science and Engineering, 1(4), pp.301-305.2010.

[8] Neelamegam and E Ramaraj.“Classification algorithm in data mining: An overview”. In: International Journal of P2P Network Trends and Technology (IJPTT) 4.8 (2013), pp. 369–374.

[9] Sahani, R., Rout, C., Badajena, J.C., Jena, A.K. and Das, H., Classification of Intrusion Detection Using Data Mining Techniques. In Progress in Computing, Analytics and Networking (pp. 753-764). Springer, Singapore.2018.

[10] Nikam, S.S., A comparative study of classification techniques in data mining algorithms. Oriental Journal of Computer Science and Technology, 8(1), pp.13-19.2015.

[11] Gorade, M., Deo, P. and Purohit, P. A Study of Some Data Mining Classification Techniques. International Research Journal of Engineering and Technology(IRJET), 4(4).2017.

[12] Zhang, G. Neural Networks For Data Mining. Data Mining and Knowledge Discovery Handbook, pp.419-444. 2009.

[13] Flach, P. and Lachiche, N. Naive Bayesian Classification of Structured Data. Machine Learning, 57(3), pp.233-269.2004.

[14] Irina, R. An empirical study of the naive Bayes classifier. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence.. [online] Available at: http://www.research.ibm.com/people/r/rish [Accessed 2 Mar. 2019].(2001).

[15] Boser, B., Guyon, I. and Vapnik, V. A training algorithm for optimal margin classifiers. Pittsburgh, PA.: ACM Press, pp.144–152.1992.

[16] Sch olkopf, B., Tsuda, K. and Vert, J .Kernel Methods in Computational Biology. MIT press. 2004.

[17] Scholkopf, B. and Smola, A. 2002 Learning with Kernels. Cambridge, MA: MIT Press. 2002.

[18] Johnson, W.G., Data Mining and Machine Learning in Education with Focus in Undergraduate CS Student Success. In Proceedings of the 2018 ACM Conference on International Computing Education Research (pp. 270-271). ACM. 2018.

[19] Khare, K., Lam, H. and Khare, A., Educational Data Mining (EDM): Researching Impact on Online Business Education. In On the Line (pp. 37-53). Springer, Cham.2018.

[20] Warren E Lacefield and E Brooks Applegate.“Data Visualization in Public Education: Longitudinal Student-, Intervention-, School-, and District Level Performance Modeling.” In: Online Submission (2018).

[21] Michael Thomas and Anouk Gelan. Special edition on language learning and learning analytics. 2018.

[22] John Fink et al. “Using data mining to explore why community college transfer studentsearn bachelor’s degrees with excess credits” .In:(2018).

[23] Poulami Dash and V. Vaidhehi. “Enhanced Elective Subject Selection for ICSE School Students using Machine Learning Algorithms”. In: Indian Journal of Science and Technology 10.21 (2017). issn: 0974 - 5645. url: http://www.indjst.org/index.php/indjst/article/view/109551.

[24] Olaniyi, A.S., Kayode, S.Y., Abiola, H.M., Tosin, S.I.T. and Babatunde, A.N., Student’s Performance Analysis Using Decision Tree Algorithms. Annals. Computer Science Series, 15(1).2017.

[25] Saurabh Pal and Vikas Chaurasia. “Performance Analysis of Students Consuming Alcohol Using Data Mining Techniques”. In: International Journal of Advance Research in Science and Engineering 6.2 (2017), pp.238– 250.

[26] Alaa M El-Halees and Ibrahim M Abu-Zaid .“Automated Usability Evaluation on University Websites using Data Mining Methods”. In: AutomatedUsabilityEvaluationonUniversityWebsitesusingDataMiningMethods 6.11 (2017).

[27] Mohammed Hussain et al. “Mining educational data for academic accreditation: aligning assessment with outcomes”. In: Global Journal of Flexible Systems Management 18.1 (2017), pp. 51–60.

[28] Mohammed M Abu Tair and Alaa M El-Halees. “Mining educational data to improve students’ performance: a case study”. In: Mining educational data to improve students’ performance: a case study 2.2 (2012).

[29] Paul Baepler and Cynthia James Murdoch.“Academic analytics and data mining in higher education”. In: International Journal for the Scholarship of Teaching and Learning 4.2 (2010), p. 17.

[30] Weakley, Jonathon JS, et al. "Visual Feedback Attenuates Mean Concentric Barbell Velocity Loss and Improves Motivation, Competitiveness, and Perceived Workload in Male Adolescent Athletes." The Journal of Strength & Conditioning Research 33.9 (2019): 2420-2425.

[31] Sewaiwar, Purva and Kamal Kant Verma. “Comparative Study of Various Decision Tree Classification Algorithm Using WEKA.” (2015).

[32] Yoan Martínez López, Julio Madera and Ireimis Leguen de Varona. “Study Of The Performance Of The K* Algorithm In International Databases.” (2016).

Climate Change Adaptation and Resilience through Big Data

Md Nazirul Islam Sarker[1], Bo Yang[2]*, Yang Lv[3], Md Enamul Huq[4], M M Kamruzzaman[5]
School of Political Science and Public Administration, Neijiang Normal University, Neijiang, China[1]
Sichuan Radio and TV University, Chengdu, China[2]
School of Public Administration, Sichuan University, Chengdu, China[2, 3]
State Key Laboratory of Information Engineering in Surveying, Mapping & Remote Sensing, Wuhan University, Wuhan, China[4]
Department of Computer and Information Science, Jouf University, Sakaka, Al-Jouf, KSA[5]

***Abstract*—The adverse effect of climate change is gradually increasing all over the world and developing countries are more sufferer. The potential of big data can be an effective tool to make an appropriate adaptation strategy and enhance the resilience of the people. This study aims to explore the potential of big data for taking proper strategy against climate change effects as well as enhance people's resilience in the face of the adverse effect of climate change. A systematic literature review has been conducted in the last ten years of existing kinds of literature. This study argues that resilience is a process of bounce back to the previous condition after facing any adverse effect. It also focuses on the integrated function of the adaptive, absorptive and transformative capacity of a social unit such as individual, community or state for facing any natural disaster. Big data technologies have the capacity to show the information regarding upcoming issues, current issues and recovery stages of the adverse effect of climate change. The findings of this study will enable policymakers and related stakeholders to take appropriate adaptation strategies for enhancing the resilience of the people of the affected areas.**

Keywords*—Disaster resilience; administrative resilience; community resilience; disaster management; environmental management

I. Introduction

Climate change is considered as a global challenge. The developing countries are facing increasing vulnerability due to global climate change. A long-term technology-driven approach is necessary to reduce vulnerability and disaster risk. Since the nature of global climate change is very complex, it requires a context-specific innovative approach to develop the adaptive, absorptive and transformative capacity of the social system. The enhanced capacities of a social unit make itself resilient against the adverse effect of climate change [1]. The adaptation strategy of vulnerable communities cannot help properly from huge direct and indirect damages of natural disasters as it is due to climate change [2]. It is now well established that technology can help to handle the situation of natural disasters. Technology can help to develop a strategy for disaster management from early warning about the disaster to post-disaster management. In every step of disaster management, policymakers, leaders, researchers, and administrators can use technology for managing the adverse effect of climate change.

The word 'resilience' originates from Latin word *resilio* that means 'to jump back'. Walker and Salt [3] have claimed that 'resilience' originated from ecological research where Holling [4] sought to differentiate between an ecological system that persists in a condition of equilibrium or stability, and response of dynamic systems when they are stressed and move from this equilibrium. A resiliency perspective is an understanding of a system's adaptive capacity [5]. Disaster management is an integrated process of preparedness, response, recovery and mitigation. All the four processes are tools of enhancing adaptive, absorptive and transformative capacity of an individual, group or community and ensure resilience. Resilience is a holistic concept that enables a social unit to bounce back from the affected condition to the previous normal condition after facing any stress caused by climate change [6]. It also focuses on an integration of adaptive, absorptive and transformative capacity. Resilience actually focuses on the multi-dimensional capacity of a social unit to face disasters and successfully manage it through reducing vulnerability and enhancing capacities [7]. Resilience also can address all the root causes of disasters through context-based adaptation strategy. It actually focuses on the way of capacity building of people so that people can reduce vulnerability, potential threat, stress, challenges and risk related to natural disasters. Big data technology can help to enhance climate resilience by context specific policy making, administration, research and leadership [8].

As a new paradigm, a big data approach is considered as the most effective method for taking quick and effective decisions [9]. It can analyze the huge amount of data obtained from various sources such as weather data, social media, electronic and print media, non-governmental organizations, voluntary community organizations, and various social networking sites. Big data provides a big opportunity for communication which can help susceptible community people about an upcoming threat, challenges, risks and disasters [10]. Communication provides a way to communicate with each other before, during and after a disaster to inform the condition to one another and make preparation and also acts as a source of big data [11]. Big data encourage researchers and policymakers to conduct an in-depth analysis of communication data from a mobile phone, social media and other communication devices which has big rationality for disaster resilience.

*Corresponding Author

Many pieces of research have already been done on climate change impacts [12,13] disaster management [14], disaster resilience [8,10], and big data application in the environmental management field [10-12] but the potential of big data for climate resilience is still lacking. Since climate change causes natural disasters, vulnerability, and scarcity of natural resources, so big data can be an effective approach to enhance climate resilience. Frequent natural disasters are also the effect of rapidly changing climatic conditions. A big data-driven effective strategy can be helpful to increase resilience. Therefore, this study aims to explore the potential of big data for enhancing climate resilience.

II. Methodology

This study is based on a qualitative approach, particularly the desk literature review. A systematic literature review has been done on the assessment of the last 10 years of literature. Recent data has been collected for participating in the ongoing debate on the potential of big data for climate resilience. This study mainly considers the big data approaches which have a potential contribution to enhancing climate resilience. A desk literature review is considered an indispensable part of developing a new paradigm of a potential field. Therefore, recent related data has been searched extensively in popular databases like the web of science, Engineering village, and Scopus. Many keywords such as ‘climate change, resilience, disaster, vulnerability, adaptive, absorptive and transformative capacity, big data, and disaster management have been used. The data collection has been done from October to November 2019. The desk review has guided by certain criteria such as (a) is this study focuses on big data for climate resilience? (b) Is this study articulate climate change adaptation using big data approaches? and (c) Is full text of this study available? Certain exclusion criteria have also been followed such as articles other than English language, duplication and article having a similar concept.

III. Results

A. Systematic Analysis Results

Qualitative document selection has been done by following the guidelines of Systematic Review and Meta-Analysis (PRISMA) [17]. PRISMA approach comprises four stages for quality document selection such as identification, screening, eligibility and included. After searching the renowned databases, this study obtained 529 documents from the main search with other 5 documents from the reference list. In the screening stage, 397 documents have been removed after the abstract screening. Similarly, in the eligibility stage, 109 documents have been removed due to the non-availability of the full text, non-relevancy, and documents not focusing on big data and climate resilience properly. Finally, most relevant 28 documents have been considered for in-depth analysis comprising journal articles, books, book chapters, and working papers (Fig. 1).

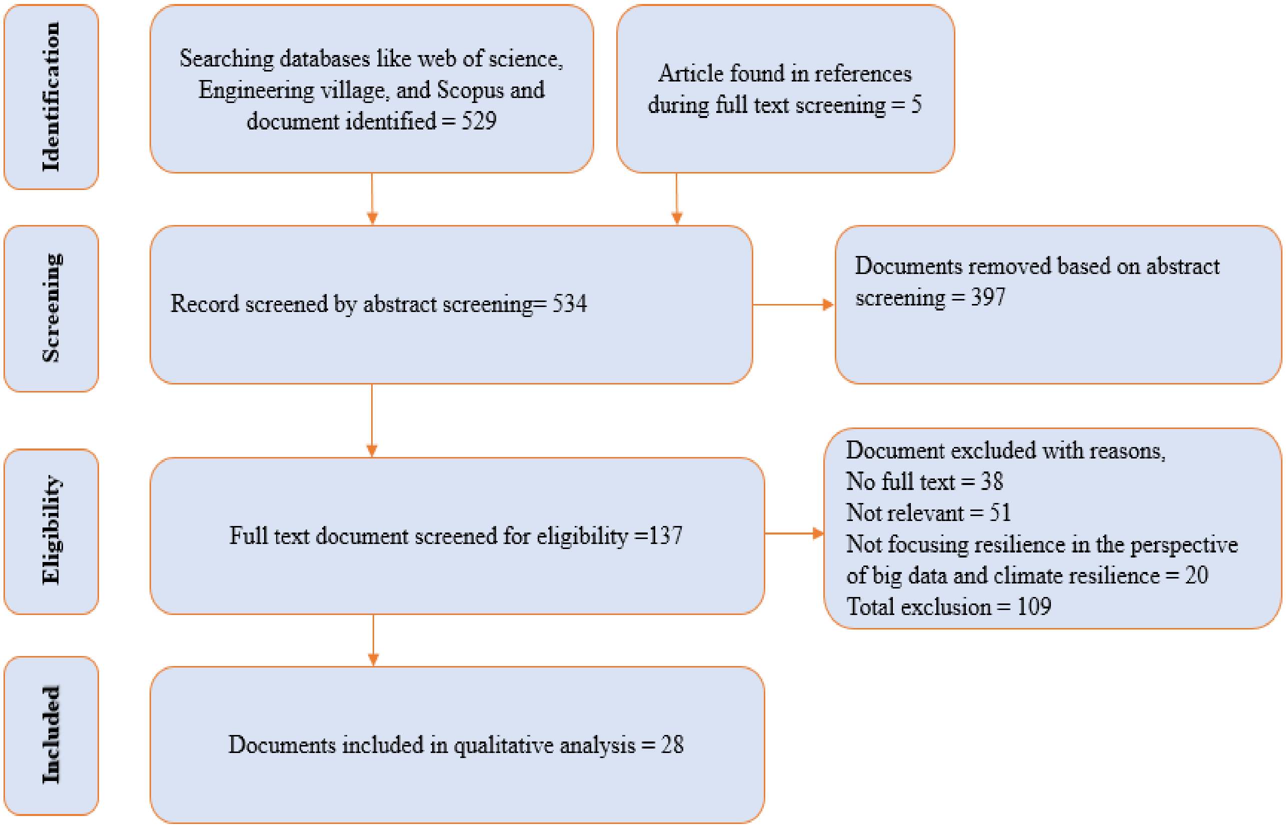

Identification
Searching databases like web of science, Engineering village, and Scopus and document identified = 529
Article found in references during full text screening = 5
Screening
Record screened by abstract screening= 534
Documents removed based on abstract screening = 397
Eligibility
Full text document screened for eligibility =137
Document excluded with reasons,
No full text = 38
Not relevant = 51
Not focusing resilience in the perspective of big data and climate resilience = 20
Total exclusion = 109
Included
Documents included in qualitative analysis = 28

Fig 1. PRISMA Flow Diagram of Document Selection.

B. Analytical Results

a) Sources of Big Data: This study is analyzed more relevant 28 documents and revealed the potential sources of big data. The major sources of big data are satellite imagery, aerial imagery and videos, wireless sensor web network, Light Detection and Ranging (LiDAR), simulation data, spatial data, crowdsourcing, social media and mobile GPS, and call record. The characteristics and recommended sources of big data are presented in Table I.

b) Recommended Phases of Disaster Management: The majority of the researchers suggested four main steps of disaster management and climate change resilience such as preparedness, mitigation, response and recovery that can directly use big data technologies for managing a disaster. This study considers all the possible climate change adaptation technologies that can help to enhance resilience of the affected people. Besides, most of the studies also suggested some sub-components of disaster management and resilience which provides an opportunity to use big data technologies at the proper time and places (Table II).

TABLE I. SOURCES OF BIG DATA

Sources of big data	Characteristics	References
Satellite imagery	High resolution, multi-technical and dimensional imagery Land use system, water bodies, direction and damaged items	Qadir et al. [18] Tomaszewski et al. [19] Park & Johnston [20]
Aerial imagery and videos	Unnamed aerial vehicles Spatial resolution of image Various sensors such as camera, infrared, ultra-violet, radiation and weather sensors	Yu et al. [16] Anbarasan et al. [21]
Wireless sensor web network	Increase response time and success delivery, reducing the latency Effective communication	Adeel et al. [22] Ogie et al. (Ogie et al. 2019) Ha [24]
Light Detection and Ranging (LiDAR)	Exact ground condition Authentic and reliable source Detect structural damages	Yu et al.[16] Shan et al. [25] Carley et al. [26]
Simulation data	Effective prediction Meteorological and land surface phenomena Agent-based modeling	Hyslop [27] Carley et al. [26]
Spatial data	Geographic information system (GIS) Vulnerability assessment and prediction	Tomaszewski et al. [19]
Crowdsourcing	Online platform Real time data	Ogie et al. [23] Clark & Guiffault [29]
Social Media	Multi-dimensional communication tool Real time data Support all phases of disaster management	Resnyansky [30] Schemp et al.[31] Enenkel et al. [35]
Mobile GPS and call record	Global Positioning System (GPS) Call detail records (CDR)	Qadir et al.[18] Gupta et al. [32]

TABLE II. PHASES OF DISASTER MANAGEMENT

Main components	Available technology	References
Preparedness	Remote sensing imagery Social media data Crowdsourced data	Lv et al. [33] Horita et al. [34]
	Remote sensing imagery (TRMM rainfall, Radarsat SAR, and Namibia Flood SensorWeb) Social media data	Enenke et al. [35] Ragini et al. [10]
Mitigation	GIS Moderate Resolution Imaging Spectroradiometer (MODIS) Crowdsourced data Mobile Metadata	Tomaszewski et al. [19] Horita et al. [34]
	Remote sensing imagery H212 model (Hurrican Weather research and Forecasting) FLOR model (Forecast Oriented Low Ocean Resolution) CYGNSS (Cyclone Global Navigation Satellite System) Airborne radar resolution	Goldenberg et al. [36] Masood et al. [37]
Response	Remote sensing imagery GEN-CAN (Global Earth Observation Catastrophe Assessment Network) Social media data	Enenkel et al. [35]
	Aerial adhoc networks SUAVs (Small Unmanned Aerial Vehicles) Team Phone	Felice et al. [38] Lu et al. [39]
Recovery	Mobile Metadata Remote sensing imagery Quick bird imagery Social media data	Contras et al. [40]

IV. DISCUSSION

A. Big Data Sources

1) Satellite imagery: Satellite imagery provides quantitative and qualitative data for disaster management which can help to conduction management operation as well as risk reduction. It can be frequently used for assessing the condition of post-disaster [18]. The major contribution of remote sensing such as high resolution, multi-technical and dimensional imagery that provides support for planning pre and post-disaster assessment. Satellite imagery provides information about changing land-use systems, water bodies, direction and damaged items of the affected area [19]. This information can help to make proper decisions about rescue methods. It is not only providing general images but also three D-dimensional images with an attitude that can easily help to detect the affected areas and level of damages [16]. It also helps to identify damaged buildings and volumes of disaster-affected areas. Satellite imagery is considered one of the major methods for disaster management due to its usage on the reduction of risk related to flood, landslide and human settlement [20].

2) Aerial imagery and videos: Unnamed Aerial Vehicles (UAVs) have been used for capturing aerial image which can be played a vital role in creating situational awareness [16]. It is considered a better method than satellite imagery due to speed and spatial resolution of the image. It can be used as an advanced level tool for detecting fine cracks, damaged structure and the extent of damages. UAVs comprises a different kind of sensors such as camera, infrared, ultra-violet, radiation and weather sensors along with spectrum analyzers [21]. It is a tool that can supply useful information to transportation planning related to real-time and situational information. UAVs are considered as an authentic data source that can help to identify real damaged caused by disasters.

3) Wireless sensor web network: Technology related to wireless sensor web (WSW) can be used for easy warning systems which helps to take preparation for saving assets from natural disasters [22]. Situational awareness can be done by using WSW [23]. Integrated use of the WSW network can enhance response time, reducing the latency as well as increasing success delivery. These technologies also ensure a connection between the affected population and the rescue team. WSW based IoT technology provides better communication in the disaster affected areas where communication structure damaged by natural disasters [24].

4) Light Detection and Ranging (LiDAR): Exact ground conditions of disaster-affected areas can be easily detected by LiDAR by using an advanced elevation model [12]. Though it is a little bit time consuming and expensive, it provides authentic and reliable information. It explores the real condition by providing high resolution [23]. The ability of LiDAR is very helpful for geological, features and mapping. It is well recognized that LiDAR can provide accurate data for water and flood assessment as well as prediction of future flooding [13]. It also provides reliable information about structural damage as well as elevation changes by natural disasters.

5) Simulation data: A simulation is a key approach for prediction. Numerical simulation can be a good approach for predicting future natural disasters by analyzing meteorological and land surface phenomena as well as different kinds of pollutions [24]. It also provides 3D modeling that can help to predict probable damage of natural disasters. Generally, huge data is generated at the time of disasters. Disaster management requires proper production, verification, validation, and improvement of data for exploring real complexity caused by natural disasters [25]. Simulation data is also helpful for assessing environmental changes through agent-based modeling. An ecological model can provide realistic information about landslide by using simulation data [13].

6) Spatial data: Spatial data is helpful for disaster management especially for vulnerability assessment and prediction of natural hazards. Tomaszewski et al. [18] conducted a study on geographic information systems (GIS) and mentioned that GIS data such as FEMA, data feeds, World Bank data, national as well as open street map is helpful for disaster management. Spatial data is frequently used for disaster resilience in the developing countries.

7) Crowdsourcing: Crowdsourcing provides an opportunity to work a large number of people in an online platform for achieving a common goal. In disaster-related crowdsourcing, many affected people can share their idea, experience, and practices for disaster management [26]. In that platform, disaster victims can share real-time information. Though it is a good source for big data, it has still some challenges especially from the credibility of the data to decision making [17]. The collection, processing, and analysis of crowdsourced data require advanced tools because of its nature and volume. Crowdsourced data is also helpful for finding out the location of the disaster-affected area [27]. It is convinced to collect crowdsourced data by using a mobile and online platform.

8) Social media: Nowadays social media is playing a vital role in almost all aspects of life. It is one of the main big data sources. Social media is considered one of the top communication tools for disaster management information [26]. It provides a piece of multi-dimensional information regarding disaster events. Though social media has short-comings due to its different kinds of data, it is still effective for disaster management [28]. Communication for disaster management was dealt with by participating organizations, victims and affected populations, vulnerable community and areas in the traditional model [29]. But big data-driven technologies can easily handle the issues very accurately and timely [41]. Since various disaster genres such as early warning, caution instruction, and immediate interaction are connected with disaster type, phases and causes, so technology-based communication can speed up the process of disaster management [30]. Various social media such as Facebook, Twitter, WhatsApp, IMO, WeChat and QQ have a great impact on almost all phases of disaster management. Social media can be used in various ways for disaster management. Social media data should be collected carefully and then process, analyze and decision making are necessary for disaster management [21]. Similarly, the decision about disaster can be easy spread-out using social media. Scholar recognized the importance of social media for disaster management phases and ensuring resilience.

9) Mobile based GPS and record of call data: The mobile phone acts as an important instrument in a disaster situation to contact family, relatives, and friends as well as to know the location for moving to a safe place. Integrated sensors of mobile phones help to identify the most affected peoples as well as the urgent needs of resources [17]. But sometimes natural disaster disrupts the electricity connection which causes an interruption of getting mobile phone services in disaster-affected places. Mobile-based GPS (Global Positioning System) is a vital way to collect mobile-based sensing data for the detection of people's behavior and movement at the time of natural disasters [18]. It helps to get real-time data of disasters regarding the human reaction to the effect of natural disasters as well as warning and evacuation process [12]. GPS helps to identify the location, altitude, magnitude and related issues of natural disasters. GPS is generally working based on three basic criteria such as proper location, comparative movement and real-time.

Call detail records (CDR) of mobile companies can record all the calls during disasters which provide a huge number of data that is real-time and need-based [42]. Disaster management personnel can easily use this big data to make quick decisions for ensuring immediate services to disaster-affected people [11]. Since the size of CDR data is large so big data approaches can take the opportunity to handle it. CDR tools can also collect the data related to human movement as well as behavior in the social network regarding natural disasters [43]. This approach can also collect the identity of the sender and receiver as well as the data of calling and SMS. From CDR data, concerned personnel can know the population density and size of the total population in the disaster-affected region [9]. Since all the subscriber are under the cellular network, it is useful to get accurate data by using this approach.

B. Big Data Approach for Climate Resilience

Vulnerability is an emerging concept across disciplines, useful in understanding and assessing the status of people's condition in the face of natural hazards. The major characteristics of climate change vulnerability are dynamic and influence people's social and biophysical processes and systems. Significant mobilization is necessary from the government, non-governmental organizations, researchers, and farmers to develop successful adaptation strategies [44]. The people of developing countries are vulnerable communities due to excessive dependency on agriculture and having a low income. However, these burdens may fuel the exploration of potential adaptive capacities of resource-poor communities [45]. The extent of people's susceptibility is increased due to the increasing vulnerability to natural hazards of almost all spheres of life, like the social, physical, human, financial, and natural dimensions [46]. Though the effect of natural hazards may be occasional, seasonal, or year-round, the extent of exposure is not the same for all communities.

1) Preparedness: The damage to climate change be can be reduced by effective detection and monitoring. Remote sensing data is usually a big source of big data that helps to detect any abnormalities of weather and disaster probability. Satellite remote sensing also can be used for the detection of the adverse effect of climate change. Some natural disasters like flooding and fire can be monitored by remote sensing imagery that helps to take proper measures for mitigation. Early detection helps to provide basic information to people so that people can prepare themselves to minimize the damages. Social media acts as a big source of big data that easily generated by people's communication [35]. In this case, people of social media share disaster-related data and act as a sensor. Social media supply real-time data which can help to identify hotspot of disaster. It also facilitates to provide information related to the probability of damages, location, duration, distance and the extent of natural disasters.

2) Mitigation: Satellite images can be easily used for detecting upcoming adverse conditions related to climate change because of its relationship with geographical position. Many scholars developed various tools and systems such as satellite-based flood mapping, and the Moderate Resolution Imaging Spectroradiometer (MODIS) for effective monitoring disaster events [16]. Vulnerability and time series analysis are also helped to assess the possible risk of natural disasters. Risk assessment can be done by using user-generated data and predict the probable events of natural disasters. It can save huge damage such as people's livelihood assets, infrastructure, health, and other basic public service systems. Crowdsourcing system is also used for risk assessment particularly to the oil industry through assessing potential exposures like the smell, and smoke [26]. Mobile Metadata and call times can be a great source of big data and helpful for decision-makers to avoid any unexpected situation related to natural disasters [25]. It is also used for risk assessment and decision making. Flood related risk can be minimized by using spatial data and decision support system.

3) Response: Remote sensing imagery is the key tool to assess the damages caused by natural disasters. The requirement of imagery depends on the technique and process. The rapid initial assessment required large scale but low-resolution data of remote sensing which can help to recommend for proper initiative for the prioritized area [31]. 3D resolution imagery is necessary for assessing the damages of roads and buildings. Damages of people's shelter and transportation networks require high-resolution remote sensing imagery. It is also used for detecting open spaces of disaster-affected areas. Manned aircraft are also used for getting good imagery based on UAV based aerial system that can be easily used for detecting disaster-affected areas. Crowdsource is a vital tool for damage assessment [23]. It works based on the data provided by the people of the disaster affected areas to the platform of crowdsourcing.

Climate change affected areas faces many problems such as lack of communication, coordination of rescue team and lack of stakeholder awareness. These problems create a barrier in the post-disaster phase and reduce resilience. Disaster usually causes damages to the local resources that also create a barrier to post-disaster management [32]. These barriers can be easily solved by using big data analytics through assessment of the condition of affected areas and maximum utilization of limited resources. Mobile networks are hampered in some affected areas which also causes a barrier for post-disaster resilience but this problem can be solved by aerial ad hoc networks that ensure the connectivity among users of affected areas [47]. Big data analytics can play a vital role in climate change adaptation through identifying hotspot for an urgent response, coordination of people to move out from risks and identifying proper response method [48].

4) Recovery: Recovery phase deals to recover the people and their assets and bringing them to a normal condition which enhances resilience. Various kinds of infrastructure are necessary for making a proper plan for quick recovery such as information related to relief distribution, confirmation of safety, coordination of volunteer activities and logistic supply [32]. The damaged communication networks are required to recover and improve rapidly. Big data can help to improve regular as well as adaptive optimization for increasing infrastructure network as an essential part of disaster recovery [34]. Communication and adaptation mechanism can be done properly by utilizing limited resources under the approaches of big data. Satellite imagery may be a vital source of big data in the post-disaster stage. Climate change affected areas can be easily detected by remote sensing imagery. Recovery evaluation is easily done by using remote sensing data.

V. Conclusion

Climate change vulnerability is considered as a common challenge all over the world. The potential of big data can be a key approach to handle the adverse effect of climate change globally. Therefore, this study focuses to develop context-specific adaptation strategy for enhancing climate change resilience by using big data technology. This study argues that resilience is a process of bounce back to the previous condition after facing any adverse effect. It also focuses on the integrated function of the adaptive, absorptive and transformative capacity of a social unit such as individual, community or state for facing any natural disaster. Big data technologies can show the information regarding upcoming issues, current issues and recovery stages of the adverse effect of climate change. It also argues that big data is a potential tool for policymakers, administrators, and related stakeholders to take necessary actions during and after disasters like an early warning system, weather forecasting, emergency evacuation, immediate responses, relief distribution, training need assessment and increasing trained individuals. For getting the maximum benefit from a big data approach for climate change resilience, this study suggests solving the related problems like challenges in data collection, analytics, infrastructure, gaps between human and technological capacity, ethical and political anomaly, poor coordination, privacy, and accuracy. This study recommends implementing proper infrastructure, technologies, tools and expertise for ensuring proper utilization of big data for climate resilience.

References

[1] R. Pandey, S. K. Jha, J. M. Alatalo, K. M. Archie, and A. K. Gupta, “Sustainable livelihood framework-based indicators for assessing climate change vulnerability and adaptation for Himalayan communities,” Ecol. Indic., vol. 79, pp. 338–346, 2017.

[2] S. A. P. Kumar, S. Bao, V. Singh, and J. Hallstrom, “Flooding disaster resilience information framework for smart and connected communities,” J. Reliab. Intell. Environ., vol. 5, no. 1, pp. 3–15. 2019.

[3] B. Walker and D. Salt, “Practicing Resilience in Different Ways,” in Resilience Practice, B. Walker and D. Salt, Eds. Washington, DC: Island Press/Center for Resource Economics, 2012, pp. 145–167.

[4] C. S. Holling, “Resilience and Stability of Ecological Systems,” Annu. Rev. Ecol. Syst., vol. 4, no. 1, pp. 1–23, Nov. 1973.

[5] FAO, Analysing Resilience for Better Targeting and Action. Food and Agriculture Organization of the United Nations, Rome, Italy, 2016.

[6] C. Folke, “Resilience: The emergence of a perspective for social–ecological systems analyses,” Glob. Environ. Chang., vol. 16, no. 3, pp. 253–267, 2006.

[7] W. N. Adger, T. P. Hughes, C. Folke, S. R. Carpenter, and J. Rockstrom, “Social-Ecological Resilience to Coastal Disasters,” Science (80)., vol. 309, no. 5737, pp. 1036–1039, Aug. 2005.

[8] C. Yang, G. Su, and J. Chen, “Using big data to enhance crisis response and disaster resilience for a smart city,” in 2017 IEEE 2nd International Conference on Big Data Analysis, ICBDA 2017, 2017, pp. 504–507.

[9] K. Amjad and K. Almustafa, “Architecture Considerations for Big Data Management,” Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 8, 2016.

[10] J. R. Ragini, P. M. R. Anand, and V. Bhaskar, “Big data analytics for disaster response and recovery through sentiment analysis,” Int. J. Inf. Manage., vol. 42, no. May, pp. 13–24, Oct. 2018.

[11] H. Maryam, M. Ali, Q. Javaid, and M. Kamran, “A Survey on Smartphones Systems for Emergency Management (SPSEM),” Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 6, pp. 301–311, 2016.

[12] L. Bizikova, P. Larkin, S. Mitchell, and R. Waldick, “An indicator set to track resilience to climate change in agriculture: A policy-maker’s perspective,” Land use policy, vol. 82, no. December 2018, pp. 444–456, 2019.

[13] W. Hein et al., “Climate change and natural disasters: Government mitigation activities and public property demand response,” Land use policy, vol. 82, no. August 2018, pp. 436–443, 2019.

[14] M. N. I. Sarker, Q. Cao, M. Wu, M. A. Hossin, G. M. Alam, and R. C. Shouse, “Vulnerability and livelihood resilience in the face of natural disaster : A critical conceptual review,” Appl. Ecol. Environ. Res., vol. 17, no. 6, pp. 12769–12785, 2019.

[15] N. Agrawal, “Disaster Resilience,” in Natural Disasters and Risk Management in Canada, N. Agrawal, Ed. Springer Science+Business Media B.V., 2018, pp. 147–191.

[16] M. Yu, C. Yang, and Y. Li, “Big data in natural disaster management: A review,” Geosciences, vol. 8, no. 5, pp. 1–26, 2018.

[17] D. Moher et al., “Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement,” PLoS Med., vol. 6, no. 7, 2009.

[18] J. Qadir, A. Ali, R. Rasool, A. Zwitter, A. Sathiaseelan, and J. Crowcroft, “Crisis analytics : big data-driven crisis response,” J. Int. Humanit. Action, vol. 1, no. 12, pp. 1–21, 2016.

[19] B. Tomaszewski, M. Judex, J. Szarzynski, C. Radestock, and L. Wirkus, “Geographic Information Systems for Disaster Response: A Review,” J. Homel. Secur. Emerg. Manag., vol. 12, no. 3, pp. 571–602, 2015.

[20] C. H. Park and E. W. Johnston, “An Event-Driven Lens for Bridging Formal Organizations and Informal Online Participation: How Policy Informatics Enables Just-in-Time Responses to Crises,” in Policy Analytics, Modelling, and Informatics, vol. 25, 2018, pp. 343–361.

[21] M. Anbarasan et al., “Detection of flood disaster system based on IoT , big data and convolutional deep neural network,” Comput. Commun., vol. 150, no. November 2019, pp. 150–157, 2020.

[22] A. Adeel et al., “A Survey on the Role of Wireless Sensor Networks and IoT in Disaster Management,” in Geological Disaster Monitoring Based on Sensor Networks, T. S. Durrani, W. Wang, and S. M. Forbes, Eds. Singapore: Springer Singapore, 2019, pp. 57–66.

[23] R. I. Ogie, R. J. Clarke, H. Forehead, and P. Perez, “Crowdsourced social media data for disaster management: Lessons from the PetaJakarta.org project,” Comput. Environ. Urban Syst., vol. 73, pp. 108–117, Jan. 2019.

[24] K. Ha, “Integrating the resources of Korean disaster management research via the Johari window,” Eval. Program Plann., vol. 77, no. June, p. 101724, 2019.

[25] S. Shan, F. Zhao, Y. Wei, and M. Liu, “Disaster management 2.0: A real-time disaster damage assessment model based on mobile social media data—A case study of Weibo (Chinese Twitter),” Saf. Sci., vol. 115, pp. 393–413, Jun. 2019.

[26] K. M. Carley, M. Malik, P. M. Landwehr, J. Pfeffer, and M. Kowalchuck, “Crowd sourcing disaster management: The complex nature of Twitter usage in Padang Indonesia,” Saf. Sci., vol. 90, no. May, pp. 48–61, Dec. 2016.

[27] M. Hyslop, “Comments on Standards in Information Security, Disaster Recovery, Business Continuity and Business Resilience,” in Critical

Information Infrastructures, no. January, Boston, MA: Springer US, 2007, pp. 94–144.

[28] R. I. I. Ogie, R. J. J. Clarke, H. Forehead, and P. Perez, "Crowdsourced social media data for disaster management: Lessons from the PetaJakarta.org project," Comput. Environ. Urban Syst., vol. 73, no. September 2018, pp. 108–117, Jan. 2019.

[29] N. Clark and F. Guiffault, "Seeing through the clouds: Processes and challenges for sharing geospatial data for disaster management in Haiti," Int. J. Disaster Risk Reduct., vol. 28, no. February, pp. 258–270, Jun. 2018.

[30] L. Resnyansky, "Social media data in the disaster context," Prometh. (United Kingdom), vol. 33, no. 2, pp. 187–212, 2015.

[31] T. Schempp, H. Zhang, A. Schmidt, M. Hong, and R. Akerkar, "A framework to integrate social media and authoritative data for disaster relief detection and distribution optimization," Int. J. Disaster Risk Reduct., vol. 39, no. April, p. 101143, Oct. 2019.

[32] A. Gupta, A. Deokar, L. Iyer, R. Sharda, and D. Schrader, "Big Data & Analytics for Societal Impact: Recent Research and Trends," Inf. Syst. Front., vol. 20, no. 2, pp. 185–194, Apr. 2018.

[33] Z. Lv, X. Li, and K. K. R. Choo, "E-government multimedia big data platform for disaster management," Multimed. Tools Appl., vol. 77, no. 8, pp. 10077–10089, 2018.

[34] F. E. A. Horita, J. P. de Albuquerque, V. Marchezini, and E. M. Mendiondo, "Bridging the gap between decision-making and emerging big data sources: An application of a model-based framework to disaster management in Brazil," Decis. Support Syst., vol. 97, pp. 12–22, May 2017.

[35] M. Enenkel, S. M. Saenz, D. S. Dookie, L. Braman, N. Obradovich, and Y. Kryvasheyeu, "Social Media Data Analysis and Feedback for Advanced Disaster Risk Management," in Social Web in Emergency and Disaster Management 2018, 2018, pp. 1–5.

[36] S. B. Goldenberg et al., "The 2012 Triply Nested, High-Resolution Operational Version of the Hurricane Weather Research and Forecasting Model (HWRF): Track and Intensity Forecast Verifications," Weather Forecast., vol. 30, no. 3, pp. 710–729, Jun. 2015.

[37] T. Masood, E. So, and D. McFarlane, "Disaster Management Operations – Big Data Analytics to Resilient Supply Networks," Dec. 2015.

[38] M. Di Felice, A. Trotta, L. Bedogni, K. R. Chowdhury, and L. Bononi, "Self-organizing aerial mesh networks for emergency communication," in 2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC), 2014, pp. 1631–1636.

[39] Z. Lu, G. Cao, and T. La Porta, "TeamPhone: Networking SmartPhones for Disaster Recovery," IEEE Trans. Mob. Comput., vol. 16, no. 12, pp. 3554–3567, Dec. 2017.

[40] D. Contreras, G. Forino, and T. Blaschke, "Measuring the progress of a recovery process after an earthquake: The case of L'aquila, Italy," Int. J. Disaster Risk Reduct., vol. 28, pp. 450–464, Jun. 2018.

[41] E. Alreshidi, "Smart Sustainable Agriculture (SSA) solution underpinned by Internet of Things (IoT) and Artificial Intelligence (AI)," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 5, pp. 93–102, 2019.

[42] A. Ahmad, R. Othman, M. Fauzan, and Q. M. Ilyas, "A semantic ontology for disaster trail management system," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 10, pp. 77–90, 2019.

[43] A. A. R. Madushanki, M. N. Halgamuge, W. A. H. S. Wirasagoda, and A. Syed, "Adoption of the Internet of Things (IoT) in agriculture and smart farming towards urban greening: A review," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 4, pp. 11–28, 2019.

[44] J. Cinnamon, S. K. Jones, and W. N. Adger, "Geoforum Evidence and future potential of mobile phone data for disease disaster management," Geoforum, vol. 75, pp. 253–264, 2016.

[45] J. R. Ragini, P. M. R. Anand, and V. Bhaskar, "International Journal of Information Management Big data analytics for disaster response and recovery through sentiment analysis," Int. J. Inf. Manage., vol. 42, no. September 2017, pp. 13–24, 2018.

[46] M. N. I. Sarker, M. Wu, G. M. Alam, and R. C. Shouse, "Livelihood resilience of riverine island dwellers in the face of natural disasters: Empirical evidence from Bangladesh," Land use policy, vol. 95, no. 2, p. 104599, Jun. 2020.

[47] V. Mali, M. Rao, and S. S. Mantha, "AHP driven GIS based emergency routing in disaster management," in Communications in Computer and Information Science, 2013, pp. 237–248.

[48] D. Kuroshima and T. Tian, "Detecting public sentiment of medicine by mining twitter data," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 10, pp. 1–5, 2019.

Enhanced Accuracy of Heart Disease Prediction using Machine Learning and Recurrent Neural Networks Ensemble Majority Voting Method

Irfan Javid[1], Ahmed Khalaf Zager Alsaedi[2], Rozaida Ghazali[3]
Faculty of Science Computer & Information Technology, Universiti Tun Hussein Onn, Malaysia[1, 3]
Department of Physics, College of Science, University of Misan, Maysan, Iraq[2]
Department of Computer Science & IT, University of Poonch,Rawalakot, AJK, Pakistan[1]

***Abstract*—To solve many problems in data science, Machine Learning (ML) techniques implicates artificial intelligence which are commonly used. The major utilization of ML is to predict the conclusion established on the extant data. Using an established dataset machine determine emulate and spread them to an unfamiliar data sets to anticipate the conclusion. A few classification algorithm's accuracy prediction is satisfactory, although other perform limited accuracy. Different ML and Deep Learning (DL) networks established on ANN have been extensively recommended for the disclosure of heart disease in antecedent researches. In this paper, we used *UCI* Heart Disease dataset to test ML techniques along with conventional methods (i.e. random forest, support vector machine, K-nearest neighbor), as well as deep learning models (i.e. long short-term-memory and gated-recurrent unit neural networks). To improve the accuracy of weak algorithms we explore voting based model by combining multiple classifiers. A provisional cogent approach was used to regulate how the ensemble technique can be enforced to improve an accuracy in the heart disease prediction. The strength of the proposed ensemble approach such as voting based model is compelling in improving the prognosis accuracy of anemic classifiers and established adequate achievement in analyze risk of heart disease. A superlative increase of 2.1% accuracy for anemic classifiers was attained with the help of an ensemble voting based model.**

Keywords—Deep learning; machine learning; heart disease; majority voting ensemble; University of California; Irvine (UCI) dataset

I. Introduction

Heart disease is particular reason of millions of worldwide death per year confer to the World heart federation Report of 2018. Stroke or CVDs are medically familiar as Heart disease (HD) along with blood pressure (BP), artery disease (AD) and debilitated heart disease by cause of diminish, blockade or reinforce capillaries that hamper the required amount of blood circulation to brain, heart, lungs and other body parts. Congestive heart failure is the most trivial kind of heart disease in all other categories of cardiovascular disease. In human body, work of blood vessels is to provide blood to the heart. Alternate, there are some other reasons of heart disease as well alike valves in the heart not supply properly and may be the reason of heart failure. Chest pain, anesthesia, jaw pain, neck ache, throat burn and back agony, cramp in upper abdomen are the most prevailing syndromes of heart disease. Withal to curtail imperil of heart disease, there are a few predominant aspects such as inhibited blood pressure, under control cholesterol and legitimate exercise. Particularly, heart disease is diagnose after angina, dilated cardiomyopathy, stroke or congestive heart failure. Thus, it is significant to pay attention to CVDs parameter and turn to doctors.

Moreover, confer to the WHO, people expire around 17.9 million every year due to CVDs which coincide to 31% of all deaths globally [1]. This provoke a demand of acquiring an economical arrangement especially capable to provide preamble appraisal of patient established on comparatively elementary medical tests that are economical to everybody. Machine learning (ML) [2] methods have drawn maximum amount of understanding in research society. As illustrate in diverse ongoing studies ML techniques have eventual offering maximum accuracy in classification as associated to alternative procedures for testimony classification. Carry out spectacular accuracy in prediction is crucial as it can edge to pertinent stability. Different machine learning techniques may varies in prediction accurateness. Therefore, it is demanding to perceive gimmick efficient of generating maximum accuracy in heart disease (HD) prediction. Prediction accuracy adept in the take up work is coordinated with earlier research studies. The uttermost practical appraisal formation approach is ML classification for the here and now along with experimental position. Three machine learning (ML) techniques have been practiced consist of random forest (RF), Support Vector Machine (SVM), k-nearest neighbor (KNN). In biomedical field like in diabetes prediction [3] [4], accomplice of diabetes and CVDs [5], reasoning of diabetes proteins [6], machine learning (ML) has already been practiced. There are the divergent conventional approach to use these fettle data to grab the latent material, but the accuracy of the conventional approach is very low, along with prolonged. So, we require contemporary technology which can backing this complex data to be appraised and grab conducive information. Deep learning (DL) algorithms have the ability to learn features from the provided training data, which outrun extracted features used in traditional machine learning algorithms. There are modernity architectures like recurrent neural network (RNN), convolutional neural network (CNN), Long-short-Term memory (LSTM) and gated recurrent unit (GRU). The extant networks confide on disease definitive approach. For classification of cardiac disease in patient modernity architectures like LSTM and GRU is applied on the extant dataset to evaluate the performance.

The Cleveland dataset from familiar *UCI* database was used to train and testing ML and DL models. It is substantiate dataset and it is extensively used for testing and training in deep learning (DL) and machine learning (ML) models [7]. The dataset consist of 303 patient records and 14 attribute features that are placed on acclaimed aspects and these features are consider to tie with risk of CVDs. We proposed a new hard voting ensemble method in this paper in which various deep learning and machine learning models are mixed and majority vote method is used to predict the result. By using this technique we can improved the overall accuracy in prediction result while aggregation of models produces collective comprehensive model.

The rest of the paper is formulated as follows. Section II, we have reviewed the earlier relevant work to the heart disease prediction and then in Section III we proposed the convoluted particulars of dataset, DL and ML techniques used and data preprocessing. Section IV shows the results produced by each model as well as the accuracy of the prediction proposed by hard voting model. Conclusion and future enhancement is outlined in Section V.

II. Review of Relevant Works

Deep learning and machine learning is advantageous for a divergent set of complications. One of the major application of these techniques is to predict the vulnerable variable from the values of autonomous variables. Even in the advanced countries one of the major reason of deaths is CVD [8]. In medical field artificial neural network (ANN) has been popularized to produce maximum accuracy [9]. The research conferred in [10] used the similar heart disease data as this study but divergent ML algorithms were enforced. Four discrete classification techniques were used which comprised Decision Tree, Naïve Bayes, Multi-layer perception and C4.5. Each of these models predict heart disease with maximum accuracy of 85.12% in the MLP classifier. Tree algorithms like J48 and Logistic model were implemented to predict CVDs also used the Cleveland HD dataset [11]. An observation of these approaches was conducted and maximum accuracy 84% was achieved with J48 algorithm.

With web base interface an application named “Intelligent Heart Disease Prediction System” was developed based on three classifier: DT, NB and ANN [12]. Several surveys conducted related to the ML utilization in Healthcare applications, especially in heart disease prognosis. The survey [13] conclude that Bayesian classification and DT surpass the others techniques like k-nearest neighbor, artificial neural network and clustering-based classification. Confer to the new study [14] by Kadi et al. has completed a pragmatic research after hands-on 149 papers proclaimed during the period from 2000-2015 for the prognosis of CVDs, DT, SVM and ANN were established to be the most periodically used ML techniques. An extreme machine learning (EML) were also implemented to predict heart disease (HD) by using UCI datasets repository and achieved highest accuracy of 80% [15]. GA and fuzzy logic (Hybrid genetic Fuzzy) approach trained and certified over similar UCI repository dataset with maximum accuracy of 86% [16].

According to [7] Raihan et al. developed an android based application to recommend a mock-up for data compilation for IHD. By practicing the P-value strategy and mobile interface they possessed 787 attributes and establish interrelationship amidst symptoms and Ischemic Heart Disease. They established a compelling correlation amidst features with P-value=0.0001 and Ischemic Heart Disease. Likewise, for scoring the symptoms statistical test chi-square, Fisher’s exact test and risk score tree are used. BP algorithm is used to extract attributes and syndromes in recent past 2018 [17]-[21].

In RNN section, LSTM consider as the determination with four important factors (forget gate (f $_g$), input gate (I $_g$), output gate (O $_g$) and cell state) have an ample usage for the image analysis along with text and audio signal analysis but is extensively usage in time series analysis, transcribed analysis, voice recognition and health testimony [22]. The major detriment of the RNN model was vanishing gradient problem, LSTM increased the input and output capability of RNN to solve these issues and it uses logical memory to learn sequence vector. To deal with CVDs data temporal features could be learn by Intelligent Healthcare Platform (IHP) established on attention module based LSTM framework [23]. Moreover, to predict CVDs 4. distinct repositories in conjunction with Cleveland dataset is used [24]. Decision Tree (DT) algorithm is the only algorithm comprises of C4.5 and Fast Tree Decision. Formerly, trained technique is established on every attributes of dataset. Later the best sample from datasets are preferred and used to train the model. This approach enhanced the prediction accuracy of the technique from 76.3% to 77.5% adopting C4.5 (average accuracy from datasets) along with enhancement in average accuracy of Fast Tree Decision from 75.48% to 78.06%.

Furthermore, to achieve highest accuracy in the prediction of CVDs distinct methods were used in contemporary research, a few classification algorithms determine CVDs with low accuracy. In contrast with traditional algorithm, hybrid method (include classification algorithms) have produce high accuracy. Our research work proposed a technique to enhance the accuracy of weak classification algorithms by linking them with rest of the classification algorithms. Thus, this technique enhanced the competence of such algorithms along with prediction accuracy for CVDs. The proposed study using ensemble majority voting techniques is done and the results are figure out. The results are compute to illustrate that aforementioned models can have adequate significant usage in medical field.

III. Experimental Result Analysis

In this paper, the main objective is to demonstrate CVDs prediction system using prior dataset. The purpose of this research is to use dataset which reflect real life data and grant the prediction system to conclude to any advanced data.

A. *Dataset Features Information*

For the experiment *UCI* Cleveland heart dataset repository has been used. The most effective 14 attributes were found amongst the 76 based on the comprehensive experiment. The

Cleveland dataset consist of most dominant 14 attributes and 303 samples. Along with 8 absolute features and 6 numeric features. Table I depicts the description of dataset.

In this dataset selected patients had age from 29 to 77. The value 0 is used to depict the female patients and value 1 is used to depict the male patients. There are 3-types of chest pain might be an indicators of heart disease. Typical angina type-1 is because of the blocked heart arteries due to decreased blood discharge to the heart muscles. The basic reason of type 1 angina is mental or emotional stress. And, second type occurs due to numerous reasons but sometime it may not be the reason of actual HD are known as Non-angina chest pain. The next feature is trestbps depicts the readings of resting blood pressure. Cholesterol level is depicted by Chol. Fasting blood sugar level is represented by Fbs. If Fbs is above 120 mg/dl then the value 0 is assigned and value 1 depicts if the Fbs is below the 120 mg/dl. Resting electrocardiography result is represented as Restecg. Maximum heart rate is represented by thalach, exercise cajoled by angina reported as 0 depicts no pain and 1 depicts pain is represented by exang, ST depression is cajoled by exercise is represented as oldpeak, Peak exercise slope ST segment is depicts by slope, number of major vessels colored by fluoroscopy is represented by ca, exercise test duration is represented by thal and the last one target is as class attribute. Class attribute value is used to distinguish the patient with heart disease and patient with no heart disease. Value 1 depicts patients with heart disease and value 0 depicts normal.

A correlation value was determined among every attributes of dataset and the target diagnosis in order to evaluate the data. Oldpeak, Exang, cp and thalach features have the highest correlated value with target feature. Table II depicts the correlated value with target attribute. This is very helpful in making an analysis against the data that is being handle with.

Furthermore, a heat map is also used to show the clear analysis of the correlation among all the attributes in Fig. 1.

Along with, a bar chart depicts in Fig. 2 gender dissemination of samples in *UCI* Cleveland dataset. The male percentage is almost 68.3% and percentage of females is 31.7% in dataset.

Moreover, histograms are devise for discrete features data visualization to depict the marginal features distribution compared for disease and not disease as represented in Fig. 3 to Fig. 8. It is observed that all the discrete features acquire normal distribution. Age vs. Thalach is shown in Fig. 9.

For the age distribution attribute, Fig. 3 represents the people with CVDs and people with no CVDs commonly. It can be viewed that maximum measurements exist between 40-52 years old. It is also realized that if age has a relation to having CVDs, then people in age range from 50-52 and 40-41 had a dominant consolidation with heart diseases.

Furthermore, to depict the possibility of any relation, Fig. 4 represents the maximal correlated discrete feature (thalach) is devise adjacent to age. It is observed that heart rate is commonly higher for the people with heart disease as compared to the people with no heart disease. Moreover, maximal heart rate decreased noted to a -ve correlated value of -0.3 as age increased. It is represented previously in Fig. 1.

TABLE. I. FEATURES DESCRIPTION OF THE CLEVELAND HEART DISEASE DATASET

S#	Features	Features Description	Data Type
1	Age	Patient age (completed in years)	Numeric
2	Sex	Gender of the patient [0= Female, 1= Male]	Binary
3	Cp	Type of chest pain classified in to four values [1- Typical angina Type, 2- Atypical angina Type, 3- Non-angina pain]	Nominal
4	Trestbps	Level of the patient blood pressure at resting mode in mm/Hg	Continuous
5	Chol	Cholesterol serum, mg/dl	Numeric
6	Fbs	Level of blood sugar on fasting (>120 mg/dl): 1 depict in case of true & 0 depict in case of false.	Binary
7	Restecg	At resting result of ECG is depict in three different values: 0 represented Normal state, 1 represented abnormality in ST-T wave, 2 having LV hypertrophy defined	Nominal
8	Thalach	Maximum rate of Heart recorded	Continuous
9	Exang	Angina-induced by exercise (1 represent 'yes' and 0 represent 'no'	Binary
10	Old peak ST	Exercise tempted ST depression comparative to rest state	Continuous
11	Slope	During peak exercise measured the ST segment in terms of slope represent in 3 values: [1. Up-sloping, 2. Flat, 3. Down-sloping]	Continuous
12	Ca	Ranges from 0-3 represent the number of vessels colored by fluoroscopy	Nominal
13	Thal	Status of the heart: [3. Normal, 6. Fixed defect, 7. Reversible defect]	Discrete
14	Target	Diagnosis represent in two categories: [0= Well, 1= possibility HD]	Binary

TABLE. II. CORRELATED VALUES WITH TARGET ATTRIBUTE ANALYSIS

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.000000	-0.098447	-0.068653	0.279351	0.213678	0.121308	-0.116211	-0.398522	0.096801	0.210013	-0.168814	0.276326	0.068001	-0.225439
sex	-0.098447	1.000000	-0.049353	-0.056769	-0.197912	0.045032	-0.058196	-0.044020	0.141664	0.096093	-0.030711	0.118261	0.210041	-0.280937
Cp	-0.068653	-0.049353	1.000000	0.047608	-0.076904	0.094444	0.044421	0.295762	-0.394280	-0.149230	0.119717	-0.181053	-0.161736	0.433798
Trestbps	0.279351	-0.056769	0.047608	1.000000	0.123174	0.177531	-0.114103	-0.046698	0.067616	0.193216	-0.121475	0.101389	0.062210	-0.144931
chol	0.213678	-0.197912	-0.076904	0.123174	1.000000	0.013294	-0.151040	-0.009940	0.067023	0.053952	-0.004038	0.070511	0.098803	-0.085239
fbs	0.121308	0.045032	0.094444	0.177531	0.013294	1.000000	-0.084189	-0.008567	0.025665	0.005747	-0.059894	0.137979	-0.032019	-0.028046
Restecg	-0.116211	-0.058196	0.044421	-0.114103	-0.151040	-0.084189	1.000000	0.044123	-0.070733	-0.058770	0.093045	-0.072042	-0.011981	0.137230
thalach	-0.398522	-0.044020	0.295762	-0.046698	-0.009940	-0.008567	0.044123	1.000000	-0.378812	-0.344187	0.386784	-0.213177	-0.096439	0.421741
Exang	0.096801	0.141664	-0.394280	0.067616	0.067023	0.025665	-0.070733	-0.378812	1.000000	0.288223	-0.257748	0.115739	0.206754	-0.436757
Oldpeak	0.210013	0.096093	-0.149230	0.193216	0.053952	0.005747	-0.058770	-0.344187	0.288223	1.000000	-0.577537	0.222682	0.210244	-0.430696
Slope	-0.168814	-0.030711	0.119717	-0.121475	-0.004038	-0.059894	0.093045	0.386784	-0.257748	-0.577537	1.000000	-0.080155	-0.104764	0.345877
Ca	0.276326	0.118261	-0.181053	0.101389	0.070511	0.137979	-0.072042	-0.213177	0.115739	0.222682	-0.080155	1.000000	0.151832	-0.391724
thal	0.068001	0.210041	-0.161736	0.062210	0.098803	-0.032019	-0.011981	-0.096439	0.206754	0.210244	-0.104764	0.151832	1.000000	-0.344029
Target	-0.225439	-0.280937	0.433798	-0.144931	-0.085239	-0.028046	0.137230	0.421741	-0.436757	-0.430696	0.345877	-0.391724	-0.344029	1.000000

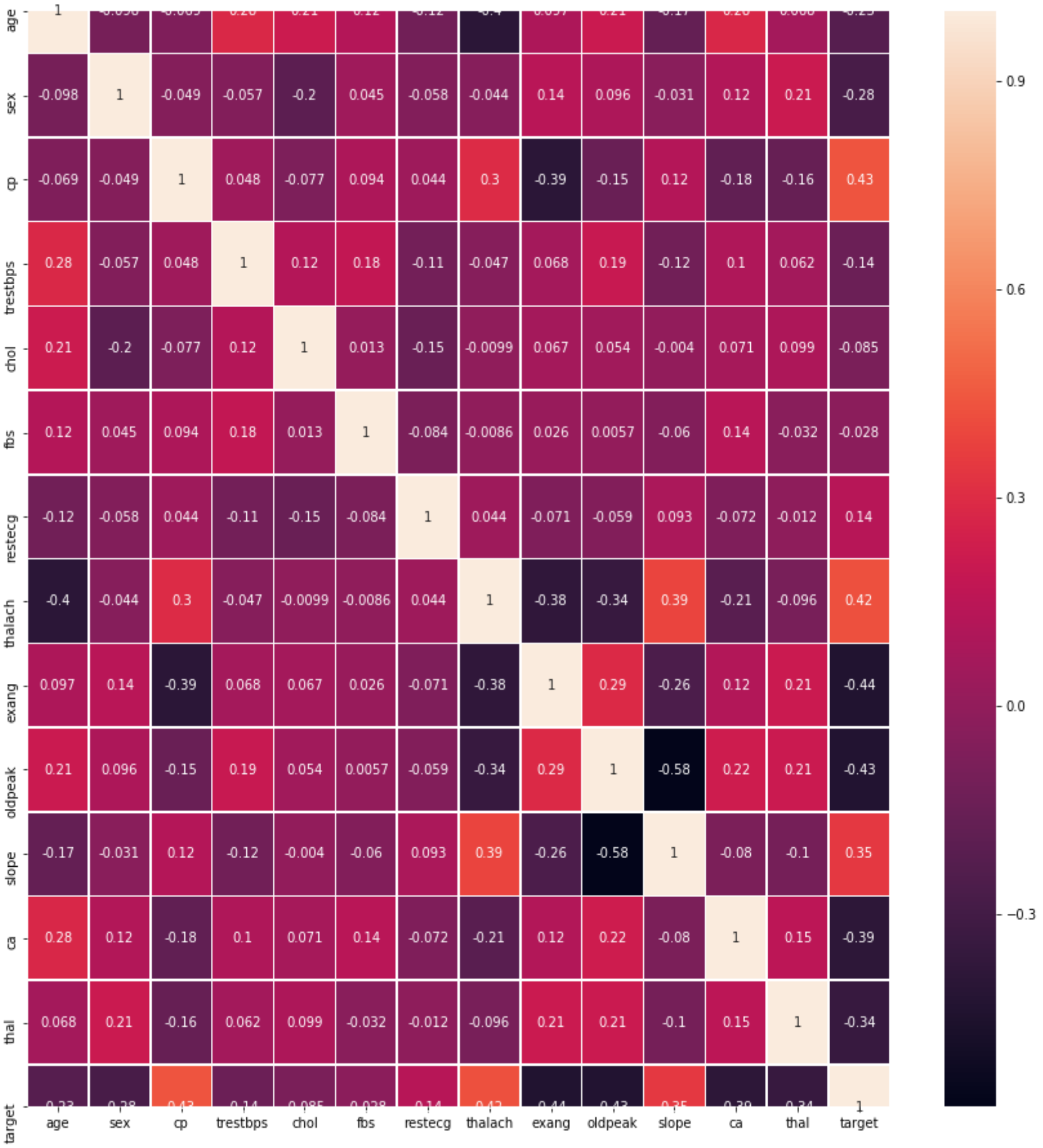


Fig. 1. Correlation-Cross Values Heat Map.

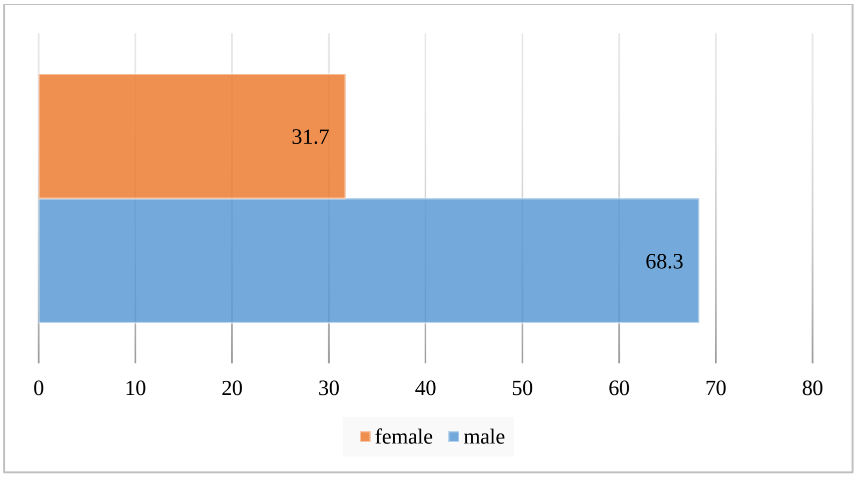


Fig. 2. Gender Dissemination

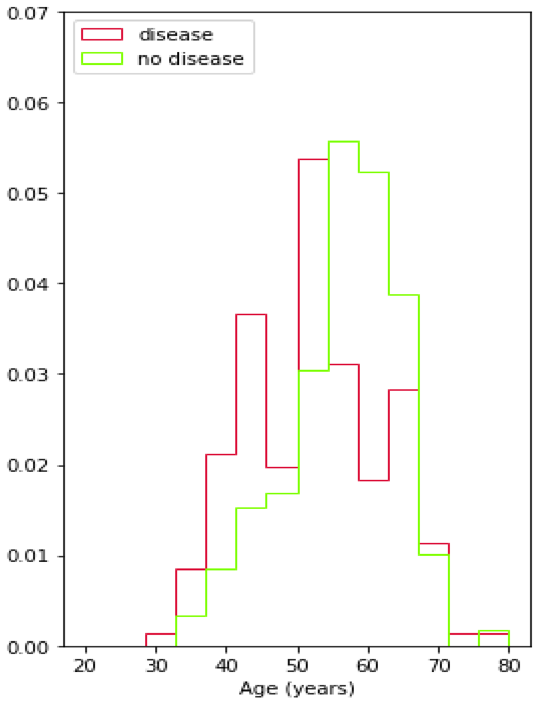


Fig. 3. Age Distribution.

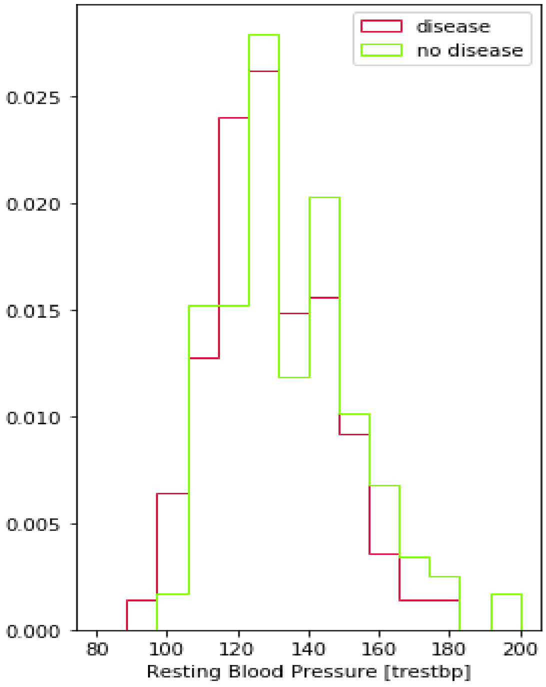


Fig. 4. Blood Pressure Distribution (at Rest).

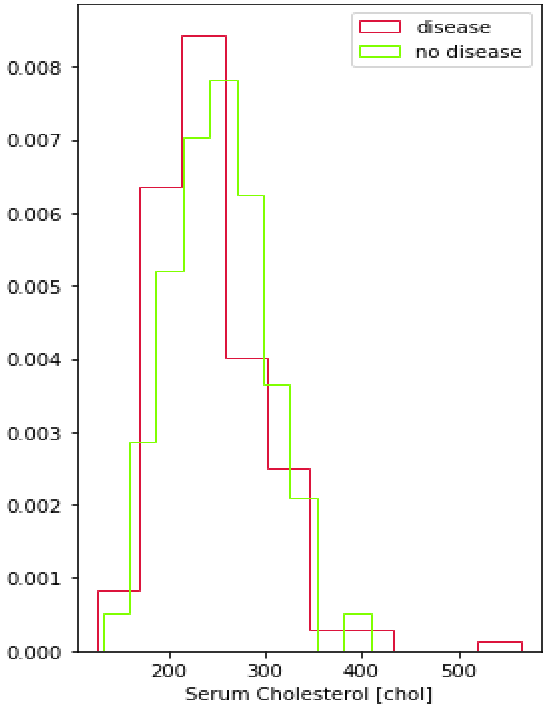


Fig. 5. Serum Cholesterol Distribution.

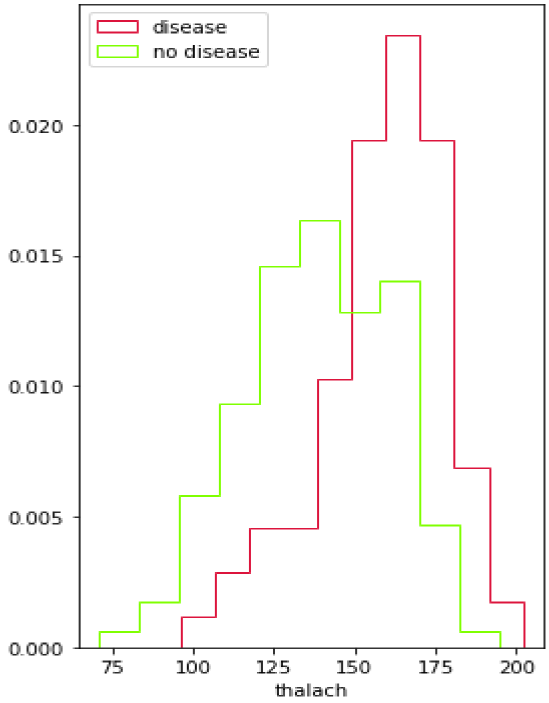


Fig. 6. Maximal Heart Rate Acquire.

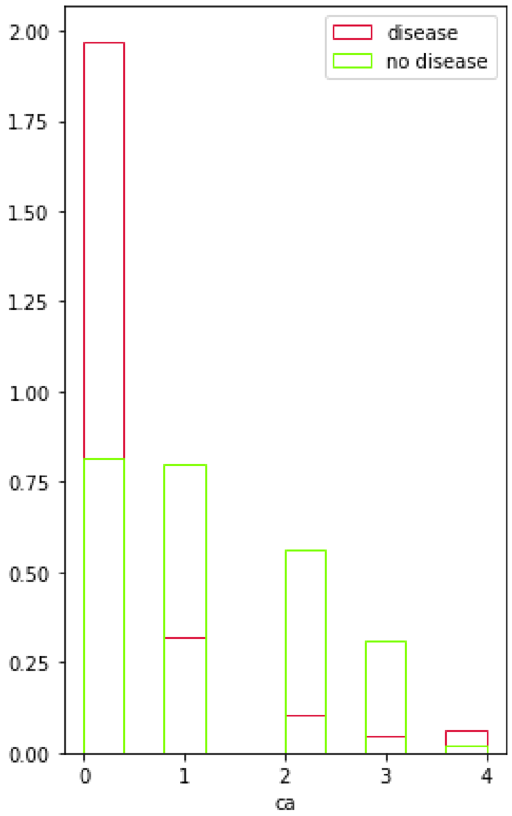


Fig. 7. Calcium Distribution.

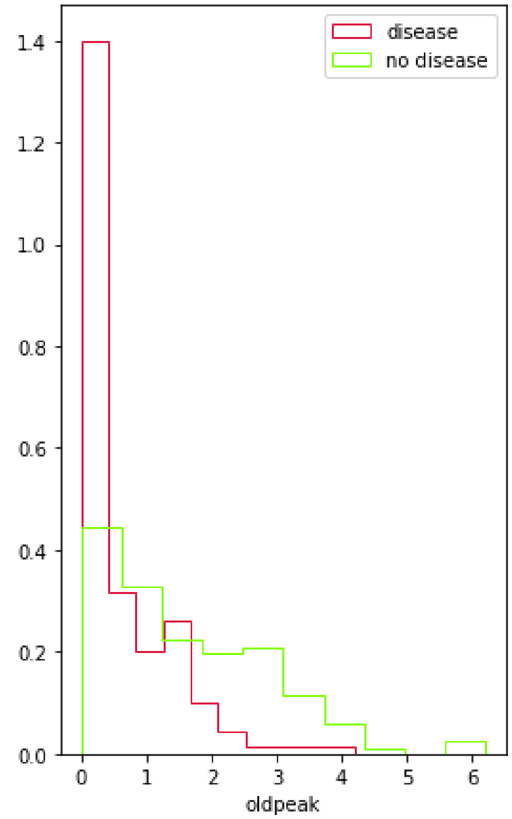


Fig. 8. Old Peak Distribution.

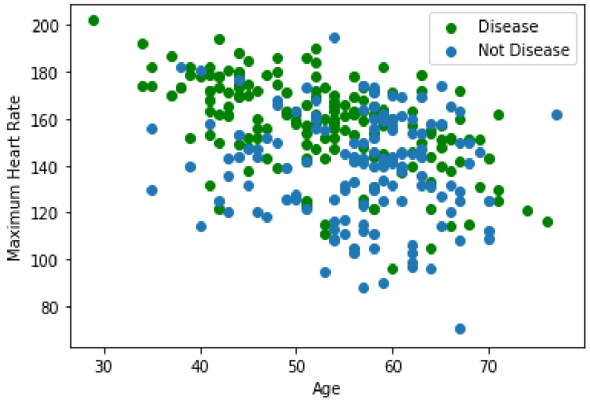


Fig. 9. Age vs. Thalach.

B. *Attribute Preprocessing*

In order to scale the maximum discrete values by using the Minimum and Maximal normalization approach, the attributes in Cleveland dataset acquire distinct proportions. As shown in eq (1), by using mentioned strategy data is transformed linearly by deducting the smallest and divide over the data range. So, the sample is categorized between zero and one which stimulate learning models to normalize the impact of distinct parameters and form a fair direction between data.

$$Z = \frac{\mathrm{N-minimum}}{\mathrm{maximum-minimum}} \qquad (1)$$

IV. Machine Learning vs. Deep Learning Models

The Cleveland heart disease dataset has been split into a testing set and train set in the scale of 80% of training set and 20% of testing data and training data set is used to train particular models. Test data is used to check the ability of a models. The working of the particular models are described in the later part.

A. *Random Forest Classifier*

It is also known as tree based classifier algorithm. Basically, name of the classifier is the indication that the algorithm build a woodland surrounded by huge number of trees. In order to get a maximum accuracy and substantial prediction, RF is an ensemble algorithm comprises on constructing numerous trees and integrate them together. This model used random samples from the training set to build set of decision trees. RFC rerun with numerous samples and compose an eventual decision established on majority voting. To handle missing information RFC is very effective but it is prone to over fitting.

B. *Support Vector Machine*

SVM was first suggested by Vladimir N. V and Alexey Ya in his study related to theory of statistical learning [25, 26]. For classification and regression purposes a supervised learning machine approach known as support vector machine (SVM) is used. In SVM a technique named trick kernel is used to revamp the information and then it identify most appropriate solution based on these alteration. At present, patient with heart disease and patient with no heart disease are classified by SVM on the basis of binary classification for k_i = +1, -1 additionally. This approach can be protected for classification in multiple classes by formulating two-multiclass classifiers [25]. A support vector machine classifier is a best approach to get reprieve hyper-plane which lie between two classes [27]. This reprieve clear hyperactive plane has numerous adequate statistical aspects. Finally, slack fickle is very informative to provide adversities of noisy data.

C. K-NN Classifier

The third classifier that was presented is the K-NN algorithm. The main purpose of this algorithm is to find the distance between the current sample along with all the trained samples, K depicts the predefined figures of adjacent points which are used for voting to the current test data's class. Certainly, classification follow established on the more classes of the K data points elected. On the bases of Grid-Search-CV more accurate results are produced and the predefined number for K in this study was selected to be 7.

D. Long-Short term Memory (LSTM)

LSTM was first proposed by Hochreiter al. is a special kind of Recurrent Neural Network (RNN) [28]. LSTM have two distinct states passed between the neurons – the cell state and the hidden layer. Cell state act as short term memory while hidden layer carry the long-term memory, commonly. There was a vanishing gradient problem with original RNN model. Therefore, RNNs are not suitable for long-term dependency data calculations. The vectors in the LSTM are added to the current node on the support of standard RNN model, which helps to solve the problems of RNN with long-term data calculations. Furthermore, LSTM model has been extensively used. LSTM layers consist of three vectors i.e., a forget vector, an input vector, and an output vector. With the passage of time many researchers proposed trivial changes to the standard LSTM model. One of the most attractive LSTM variant "peephole-connections" was introduced by Gers et al. [29]. There are numerous adaptations with small changes regarding the gated structure in the LSTM units. Here we will consider the one proposed by Graves et al. [32].

$$i_t = \sigma\ (W_{xi}\, x_t + W_{hi}\, h_{t-1} + W_{ci}\, c_{t-1} + b_i\ ) \tag{2}$$

$$r_t = \sigma\ (W_{xr}\, x_t + W_{hr}\, h_{t-1} + W_{cr}\, c_{t-1} + b_r\ ) \tag{3}$$

$$c_t = r_t\ \iota\ c_{t-1} + i_t\ \iota\ \tanh(W_{xr}\, x_t + W_{xc}\, x_t + W_{hc}\, h_{t-1} + b_c\ ) \tag{4}$$

$$o_t = \sigma\ (W_{xo}\, x_t + W_{ho}\, h_{t-1} + W_{co}\, c_{t-1} + b_o\ ) \tag{5}$$

$$h_t = o_t\ \iota\ \tanh(c_t) \tag{6}$$

where ι represent element wise product and r, o, i are the forget vector, output vector and input vector respectively. It is observed that the gating structure regulates how the new input and previous hidden state value must be unite to produce the new hidden state value.

The most attractive variant of LSTM is gated-recurrent unit (GRU) was introduced by Chung et al. [30]. The idea was to combine forget vector and input vector as single update vector. In GRU, cell state and hidden state are also merges and make some numerous changes as well. The GRU support the long term sequences and also carry the long-term memories. Therefore, proposed GRU architecture is simpler and most attractive than the original LSTM model.

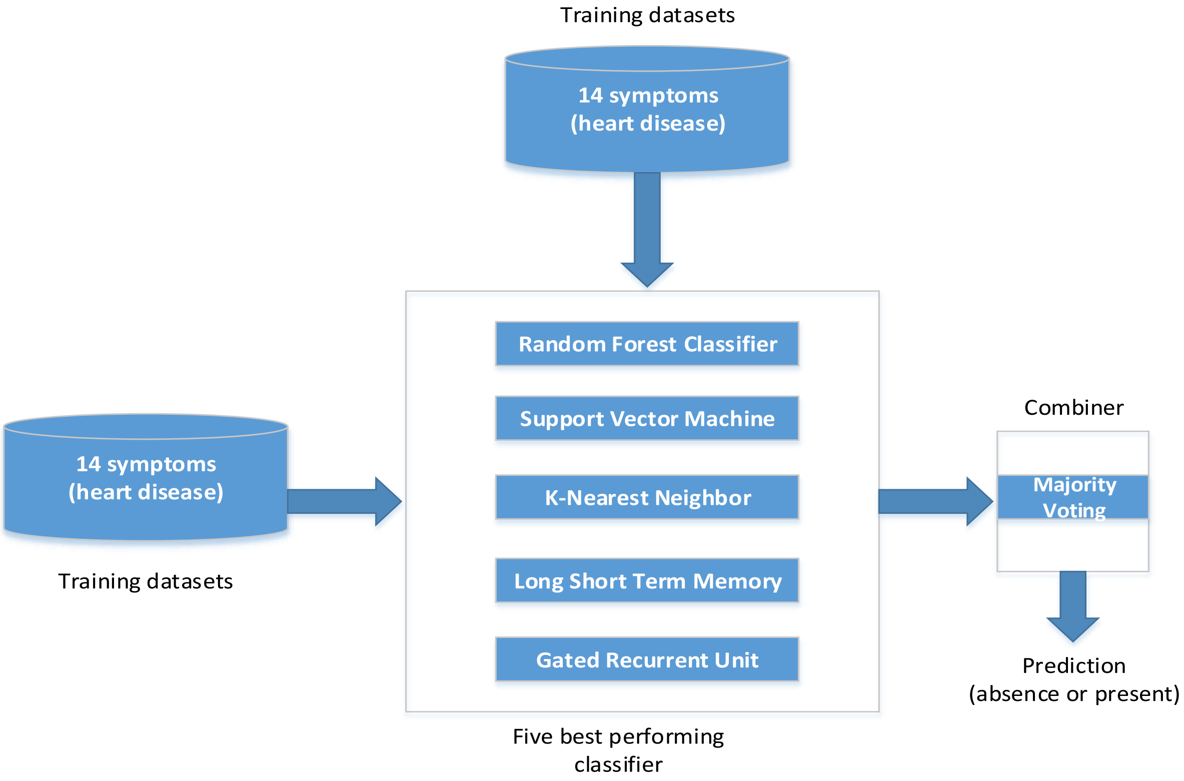


Fig. 10. Workflow of the Proposed Ensemble Vote based Model.

E. Gated Recurrent Unit (GRU)

Cho et al. [30] proposed another gating structure known as GRU (gated recurrent unit) with the purpose to carry long-term dependencies from the calculations within the GRU neuron to produce the hidden state. GRU have only one hidden state conveyed between time steps. Following are the equations determined by Chung et al. [31].

$$r_{t\,=}\,\sigma\,(W_r\,x_t + U_r\,h_{t-1}\,) \quad (7)$$

$$\tilde{h}_t = \tanh\,(W\,x_t + U(r_t \,\square\, h_{t-1})) \quad (8)$$

$$z_t = \sigma\,(W_z\,x_t + U_z\,h_{t-1}\,) \quad (9)$$

$$h_t = (1\text{-}\,z_t\,)\,h_{t-1} + z_t\,\tilde{h}_t \quad (10)$$

Where r and z are commonly the reset and update gates. It can be observed that, GRU is most simple than LSTM, and performance is far better in different experiments. In [31], Chung et al. provide a comparison related to the performance of original RNN, LSTM and GRU, using numerous datasets. It was observed that gated recurrent unit surpass the other techniques in different situations.

F. Ensemble Classifier

At the end, five models aforementioned are unite in an ensemble method where hard voting (majority vote of the models) technique is used for classification. The voting is based on the prediction of each model about each sample and final prediction is based on the majority votes, one that obtains more than 50% of the votes.

The independent classifiers output is united and plays an important role in the final output prediction of an ensemble system. As shown in the Fig.10. Therefore, one of the interesting research study is combination of classifiers in ensemble system. Majority voting approach is extensively used method for labeling the output [33]. In case of discrete outputs, like linear combination, a maximum, minimum, average or any other alternate like derriere possibilities may be used. Many times a classifier may be used as a meta-classifier for uniting outputs of ensemble-members. Due to better performance of majority voting approach over other linear and meta-classifiers has been applied in this work. Therefore, majority voting rule lies in 3 categories: (1) Unanimous-Voting method, here every models must acknowledge the prediction, (2) simple majority method, here prediction required to be partially higher than 50% of classifiers, and (3) majority voting method, here maximum figures of votes is required for the ensemble-decision. If the output of the individual classifier is independent than the majority voting rule combiner constantly enhance the prediction accuracy [34]. Suppose that a class define outputs of classifier O_i are shown as d-dimensional binary vectors:

$$[O_{i,\,1\,\ldots}\,O_{i,\,d]}\;Є\;\{0,1\}^{\,d}, \; i=1,\ldots\ldots\ldots,N \quad (11)$$

Where $O_{i,1}$=1, if classifier O_i label y in w_j , and 0 differently. The majority voting method would provide an ensemble decision for class w_k , if the below equation is satisfied:

$$\sum_{i=1}^{N}\;O_{i,\,k} = \max{}^{c}\,j{=}1\,\sum_{i=1}^{N}\;O_{i,\,j} \quad (12)$$

If we have 2 classes (c=2), the majority voting method correspond with simple majority approach (50% of vote +1). According to the equation (4) majority voting approach would predict an accurate class define at least [N/2+1] classifiers correctly predict the define class [35]. In our proposed research work, *N* = 5, it observes that our proposed approach would be able to predict correctly if more than half (at least 3 classifiers) predict the define class correctly.

V. Experimental Results and Discussion

The first classifier Random forest, to study unseen data prediction was run on the test dataset so that the approach has never overcome. Default parameters of the approach are used to run the early test and composed an accuracy of 83.6%. Along with, attributes importance was calculated in this approach and most important attributes were (ca, thalach, oldpeak). Confusion matrix obtained from this approach is shown in Fig. 11.

Also, the second classifier was Support Vector Machine (SVM) algorithm. To run the unseen test dataset the approach was developed with the default parameters. The prediction accuracy of this model was 81.31%. In Fig. 12, confusion matrix obtained from this classifier is depicted.

The third approach, known as K-Nearest Neighbor model. To run the unseen test dataset using default parameters we developed the model. The prediction accuracy get out to be 82.8%. Fig. 13 depicts confusion matrix obtained from this algorithm.

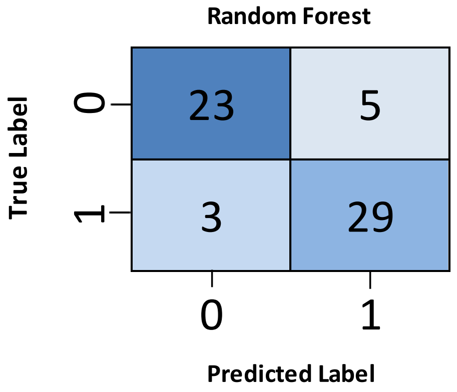


Fig. 11. Random Forest Model Confusion-Matrix.

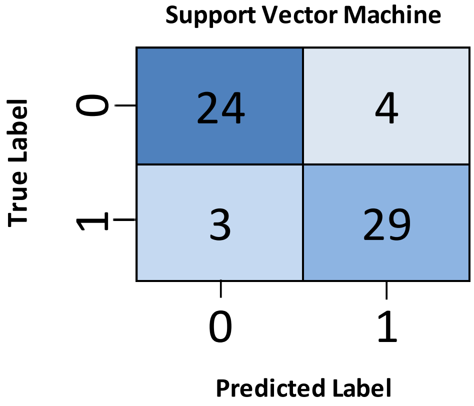


Fig. 12. SVM Model Confusion-Matrix.

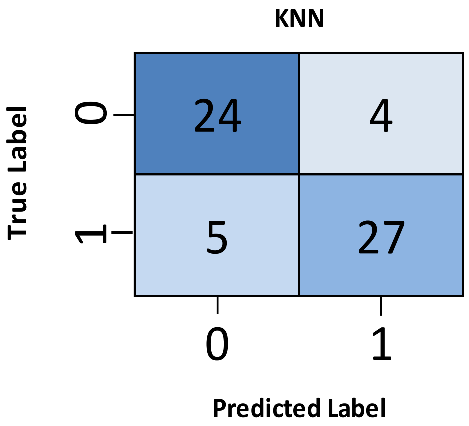


Fig. 13. K-Nearest Neighbor Model Confusion-Matrix.

Additionally, fourth approach that was developed known as LSTM model. Using the default parameters this approach was developed and classification established based on the hidden data test set. The prediction accuracy get out to be 81.31%. In Fig. 14, Confusion matrix obtained from this model is depicted.

Finally, the fifth approach that was developed was the GRU model. Using the default parameters this approach was developed and classification established based on the hidden data test set. The prediction accuracy get out to be 81.46%. Fig. 15 depicts the confusion matrix obtained from this model.

We have noticed that Random Forest and K-NN are constantly provide better prediction accuracy as compared to other classification models. The performance of the each model in accuracy prediction of Heart-Disease as shown in the Fig. 16.

Certainly, the overall prediction accuracy of this study after organizing the Hard Voting ensemble-method get out to be 85.71% which is treated a fairly required accuracy that can be further developed upon in future.

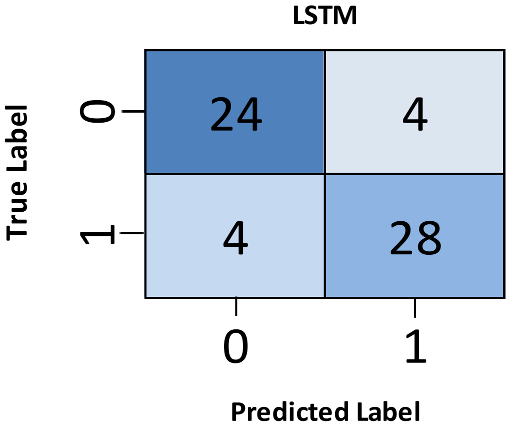


Fig. 14. LSTM Model Confusion-Matrix.

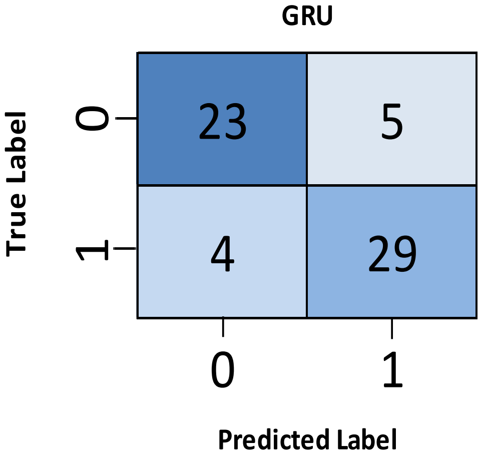


Fig. 15. GRU Model Confusion-Matrix.

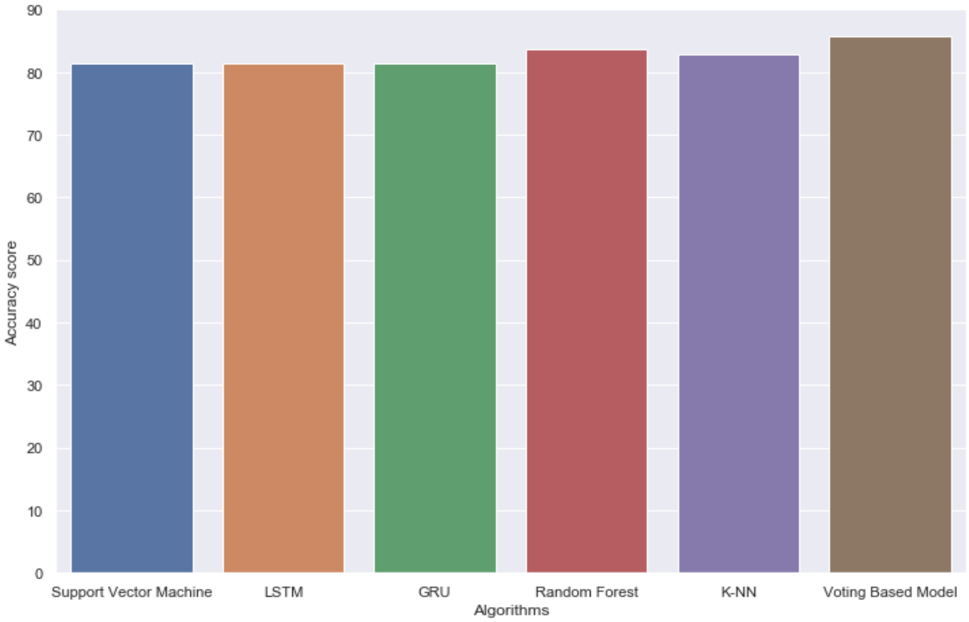


Fig. 16. The Performance Study of different ML and DL Models.

VI. Conclusion

To save the life of the human beings, early prediction of heart disease plays significant role. Here, in this paper we presented a ML and DL ensemble models that united multiple ML and DL models in order to give a maximum accuracy and vigorous model for the prediction of any possibility of having heart disease. Table III depicts the prediction accuracy comparison of Machine learning techniques (i.e. RF, SVM and KNN), deep learning models (i.e. LSTM and GRU) and proposed methodology. This Ensemble approach retained 85.71% accuracy, which surpass the prediction accuracy of every particular model. This approach may be very useful to assist the doctors to investigate the patient cases in order to legitimize their prescription. The future work of this study can be performed with different mixtures of ML and DL models to better prediction.

TABLE. III. Prediction Accuracy Comparison of the Models

Model Name	Accuracy
Random Forest	83.6%
Support Vector Machine	81.3%
K-NN	82.8%
LSTM	81.31%
GRU	81.46%
Hard Voting Ensemble Model	85.71%

References

[1] “Cardiovascular diseases (CVDs),” World Health Organization, 26-Sep2018.[Online]. Available: https://www.who.int/cardiovascular_ diseases /en/. [Accessed: 27-Apr-2019].

[2] Bache K, Lichman M (2013) UCI machine learning repository [http://archive.ics.uci.edu/ml]. University of California, School of Information and Computer Science. Irvine, CA.

[3] Bansal A, Agarwal R, Sharma R (2015) Determining diabetes using iris recognition system. Int J Diabetes Dev Ctries 35(4):432–438.

[4] Kalaiselvi C, Nasira GM (2015) Classification and prediction ofheart disease from diabetes patients using hybrid particle swarm optimization and library support vector machine algorithm. Int J Comput Algorithm 4(1):2278–2397.

[5] Bhramaramba R, Allam AR, Kumar VV, Sridhar G (2011) Application of data mining techniques on diabetes related proteins. Int J Diabetes Dev Ctries 31(1):22–25.

[6] King RD (1992) Statlog databases. Department of Statistics and Modelling Science, University of Strathclyde, Glasgow.

[7] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[8] Vanisree K, JyothiSingaraju. Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks. Int J Comput Appl April 2011;19(6). (0975 8887).

[9] L. Baccour, ``Amended fused TOPSIS-VIKOR for classication (ATOVIC) applied to some UCI data sets,'' Expert Syst. Appl., vol. 99, pp. 115125, Jun. 2018. doi: 10.1016/j.eswa.2018.01.025.

[10] D. Chaki, A. Das, and M. Zaber, “A comparison of three discrete methods for classification of heart disease data,” Bangladesh Journal of Scientific and Industrial Research, vol. 50, no. 4, pp. 293–296, 2015.

[11] R. G. Saboji, “A scalable solution for heart disease prediction using classification mining technique,” 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017.

[12] S. Palaniappan, R. Awang, in: Intelligent heart disease prediction system using data mining techniques, 2008 IEEE/ACS International Conference on Computer Systems and Applications, 2008, pp. 108–115.

[13] J. Soni, U. Ansari, D. Sharma, S. Soni, Predictive data mining for medical diagnosis: an overview of heart disease prediction, Int. J. Comput. Appl. 17 (8) (2011) 43–48.

[14] I. Kadi, A. Idri, J.L. Fernandez-Aleman, Knowledge discovery in cardiology: a systematic literature review, Int. J. Med. Inform. 97 (2017) 12–32.

[15] S. Ismaeel, A. Miri, D. Chourishi, in: Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis, IEEE Canada International Humanitarian Technology Conference, 2015, pp. 1–3.

[16] M. Raihan, S. Mondal, A. More, and M. Sagor et al., "Smartphonebased ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design," in International Conference on Computer and Information Technology (ICCIT), pp. 299-303, 2016.

[17] G. Shanmugasundaram, V. M. Selvam, R. Saravanan, and S. Balaji, "An Investigation of Heart Disease Prediction Techniques," in IEEE International Conference on System, Computation, Automation and Networking (ICSCA), pp. 1-6, 2018.

[18] P. Umasankar and V. Thiagarasu, "Decision Support System for Heart Disease Diagnosis Using Interval Vague Set and Fuzzy Association Rule Mining," in International Conference on Devices, Circuits and Systems (ICDCS), pp. 223-227, 2018.

[19] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, and V. Mareeswari, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms," in International Conference on Current Trends towards Converging Technologies (ICCTCT), pp. 1-7, 2018.

[20] I. K. A. Enriko, M. Suryanegara, and D. Gunawan, "Heart Disease Diagnosis System with k-Nearest Neighbors Method Using Real Clinical Medical Records," in International Conference on Frontiers of Educational Technologies, pp. 127-131, 2018.

[21] H. Kahtan, K. Z. Zamli, W. N. A. W. A. Fatthi, and A. Abdullah et al., "Heart Disease Diagnosis System Using Fuzzy Logic," in International Conference on Software and Computer Applications, pp. 297-301, 2018.

[22] T. Ergen and S. S. Kozat, "Online Training of LSTM Networks in Distributed Systems for Variable Length Data Sequences," IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 10, pp. 5159-5165, 2018.

[23] C. Lin, Y. Zhangy, J. Ivy, and M. Capan et al., "Early Diagnosis and Prediction of Sepsis Shock by Combining Static and Dynamic Information Using Convolutional-LSTM," in IEEE International Conference on Healthcare Informatics (ICHI), pp. 219-228, 2018.

[24] El-Bialy, R., Salamay, M., Karam, O. and Khalifa, M. (2015). Feature Analysis of Coronary Artery Heart Disease Data Sets. Procedia Computer Science, 65, pp.459-468.

[25] Vapnik VN, Vapnik V (1998) Statistical learning theory, vol 2.Wiley, New York.

[26] Vapnik V (2000) The nature of statistical learning theory.Springer, Berlin

[27] Burges CJ (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2(2):121–167.

[28] S. Hochreiter and J. Schmidhuber, ``Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 17351780, 1997.

[29] F. A. Gers and J. Schmidhuber, ``Recurrent nets that time and count," in Proc. IEEE-INNS-ENNS Int. Joint Conf. Neural Netw. (IJCNN), vol. 3. Jul. 2000, pp. 189194.

[30] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. (2014). "Empirical evalua-tion of gated recurrent neural networks on sequence modeling." [Online]. Available: https://arxiv.org/abs/1412.3555.

[31] Chung J, Gulcehre C, Cho K, Bengio Y. arXiv preprint arXiv:1412.3555.

[32] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks. In ICASSP2013, pages 66456649. IEEE, 2013

[33] J. Kittler, M. Hatef, R.P. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (3) (1998) 226–239.

[34] J. Kittler, M. Hatef, R.P. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (3) (1998) 226–239.

[35] L.I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, John Wiley & Sons, New York, 2004.

Temporal Analysis of GDOP to Quantify the Benefits of GPS and GLONASS Combination on Satellite Geometry

Claudio Meneghini[1], Claudio Parente[2]
Department of Sciences and Technologies
University of Naples “Parthenope”
Naples, Italy

***Abstract*—Global Navigation Satellite Systems (GNSS) have developed rapidly over the last few years. At present, there are GNSS receivers that combine satellites from two or more different constellations. The geometry of the satellites in relation to the receiver location, i.e. how nearly or distantly they are disposed in the sky, impacts on the quality of the survey, which is essential to achieve the highest level of position accuracy. A dimensionless number identified as Geometric Dilution of Precision (GDOP) is used to represent the efficiency of the satellite distribution and can be easy calculated for each location and time using satellite ephemeris. This paper quantifies the influence of multi-GNSS constellation, in particular GPS (Global Positioning System) and GLONASS (Globalnaya Navigazionnaya Sputnikovaya Sistema) combination, on satellite geometry considering a precise period. A new index named Temporal Variability of Geometric Dilution of Precision (TVGDOP) is proposed and analyzed in different scenarios (different cut-off angles as well as real obstacles such as terrain morphology and buildings). The new index is calculated for each of the two satellite systems (GPS and GLONASS) as well as for their integration. The TVGDOP values enable the three cases to be compared and permit to quantify the benefits of GNSS integration on satellite geometry. The results confirm the efficiency of the proposed index to highlight the better performance of combination GPS+GLONASS especially in presence of obstacles.**

Keywords—GDOP* (*Geometric Dilution of Precision); GPS (Global Positioning System); GLONASS (Globalnaya Navigazionnaya Sputnikovaya Sistema); Multi-GNSS (Global Navigation Satellite System) Constellation

I. Introduction

Satellite constellations were designed to provide three-dimensional navigation by determining the position of a receiver on land, at sea, or in space [1]. Errors in determining the position of receivers are caused by the precision that characterizes the measure of the distance to each satellite, and the placement of the satellites relatively to the receiver location, i.e. how nearly or distantly they are disposed in the sky [2].

Various techniques can be used to limit distance errors, but satellite geometry is essential to achieve the highest level of position accuracy. This geometry is often expressed using a numerical measure named “Dilution of Precision”, or DOP. There are several types of DOPs, which are all functions of the receiver-transmitter geometry. The Geometric Dilution of Precision (GDOP) is a dimensionless quantity that represents the efficiency of the satellite distribution. If GDOP value is small, the position accuracy is high. With four satellites, which is the minimum number required for a single constellation, the best geometry is achieved when all four satellites together form a tetrahedron structure in which one of them is at the zenith and the others form an equilateral triangle. The value of the GDOP is small (so the accuracy is high), if the volume of the tetrahedron is large, as well as if the number of satellites is high [3]. The number of visible satellites influences positioning accuracy, availability and reliability: considering only one system such as GPS (Global Positioning System),, the visible satellites are often scarce in areas such as mountains, open-pit mines and urban canyons [4]. Sometimes satellite geometry can be inadequate even when four or more satellites of the same system are available. In such situations, the GNSS (Global Navigation Satellite Systems) multi-constellation approach is useful to increment the number of visible satellite, perform their geometry, and improve the continuity and reliability of the positioning.

Several studies and applications have demonstrated the benefits of multi-constellation operations in a receiver. Integrated GNSS significantly improves the results compared with them by each constellation, especially in obstructed areas [5] [6] [7] [8]. For example, combining GLONASS (Globalnaya Navigazionnaya Sputnikovaya Sistema) and GPS constellations, permits to achieve accessibility to a greater number of satellites in urban canyons, and better accuracy in zones of minimal availability [9].

Today more than 70 satellites are already in view, and about 120 satellites will be available once all four systems (BeiDou, Galileo, GLONASS and GPS) are fully deployed in the next few years [10].

The Global Positioning System (GPS) was the first global positioning system in the world and was created in 1973 by the US Department of Defense for military purposes. The first GPS satellite was launched in 1977, and the global positioning service reached full operational capacity 18 years later. Since 2000, GPS has been available free of charge for civilian purposes. At the time of writing, there are 31 operational satellites in the GPS constellation [11].

Developed by Russian Federal Space Agency in 1970, the GLONASS (Global'naya Navigatsionnaya Sputnikowaya Sistema) system was created for military purposes, and is managed by the Russian Ministry of Defense. The first GLONASS satellite was launched in 1984, and the GLONASS constellation reached the full working conditions of 24 satellites in 1996. Due to a lack of funds, the number of available satellites significantly decreased until December 2011 when the constellation again increased to 24 operating satellites. In 2006, GLONASS became free of charge for public usage. Currently there are 22 active satellites [12].

We performed a temporal analysis of the GDOP in order to evaluate the impacts of GPS and GLONASS combination on the quality of the GNSS survey. The GDOP was calculated as function of date and time considering three different constellations (GPS, GLONASS and GPS+GLONASS) and a period of 32 days. The variability of the cut-off angle as well as the presence of real obstacles were also considered. To facilitate a comparison of the various constellations and situations, a new index named TVGDOP (Temporal Variability of Geometric Dilution of Precision) produced as the inverse value of GDOP is introduced.

This paper is organized as follows. Section 2 presents the concepts of satellite geometry and Dilution of Precision (DOP). Section 3 describes and discusses the adopted approach. Finally, Section 4 draws some conclusions.

II. Theoretical Background

A. Satellite Geometry Impact on Positioning

The geometric positions of the satellites in relation to the receiver can decrease navigation and survey accuracy. To explain this concept, it is worth looking at the case of two satellites. If the satellites are considered as the center, two arcs can be delineated from the satellites to represent positional lines. The first is internal with a radius equal to the true range, while the second is external with a pseudo-range as the radius.

The possible user position is defined by the intersection zone of the two arcs:

- If the distance between the satellites is significant (Fig. 1), the connection area is small. Consequently, a low position uncertainty means a good satellite geometry.
- If the satellites are near to each other (Fig. 2), the joint area is large. In this circumstance, considerable indecision regarding the position denotes a bad satellite geometry.

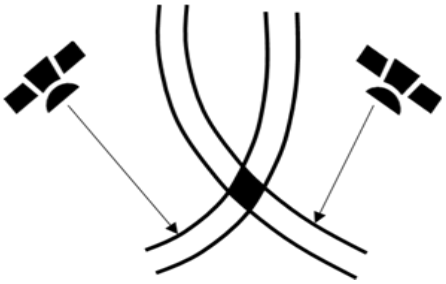


Fig. 1. A Low Position uncertainty for Two Satellites.

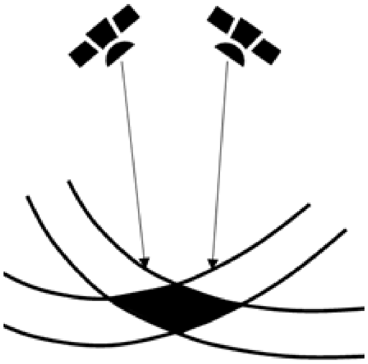


Fig. 2. A High uncertainty of Position for Two Satellites.

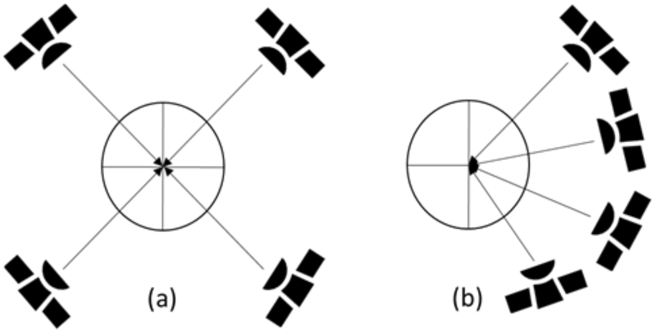


Fig. 3. Good (a) and Poor (b) Satellite Geometry.

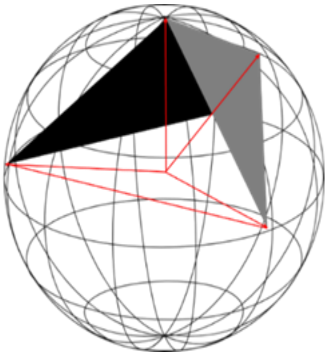


Fig. 4. The Tetrahedron Shaped by the Receiver and Four Satellites.

To fix the user position, at least four satellites (from the same constellation) are required. In this case, to gain a good geometry, all the available satellites must be distant from each other in the space. Fig. 3 represents a good and poor satellite geometry, respectively, with four satellites.

The best geometry with four satellites can be achieved if one of them is at the zenith and the others form an equilateral triangle. All the satellites together delimit a tetrahedron (Fig. 4).

If there are five or more satellites in view, only four of them, the ones corresponding to the best combination, are usually considered for redundancy reduction. To fix the receiver location, precise range measurements are necessary, more precisely, the measurements of ranges between the satellites and the receiver antenna. With four or more pseudo-range observations, it is possible to calculate both the receiver's three-dimensional coordinates and its clock offset.

The quality of the pseudo-range measurements defines the accuracy of the receiver coordinates. According to the literature [13] [14], the well-known equation to calculate the pseudo-range is:

$$P = \rho + c\,(dT - \,dt) + d_{ion} + d_{trop} + e \qquad (1)$$

where:

- P represents the pseudo-range measurement;
- ρ indicates the distance between the satellite antenna and the receiver antenna;
- c is the speed of light in a vacuum;
- dT is the receiver clock bias from the GNSS time;
- dt is the satellite clock bias from the GNSS time;
- d_{ion} is the ionospheric propagation delay;
- d_{trop} is the tropospheric propagation delay;
- e is the measurement noise in addition to multipath effect.

Considering that satellite clock bias, tropospheric delay and ionospheric delay are known, the pseudo-range equation becomes:

$$P = \rho + c\,dT + e \qquad (2)$$

A receiver must resolve a number of equations equal to the number of measurements deriving from the visible satellites. At least four observations are needed to calculate the receiver coordinates.

The equations are not linear, but can be linearized considering the original estimates for the station's position. By correcting the initial estimates, it is possible to evaluate both the receiver's current coordinates and the clock offset. In addition, the equations can be grouped and represented in a matrix form:

$$\delta P = A\delta U \,+\, n \qquad (3)$$

where:

- A is a matrix where each term defines the direction cosine vector from the receiver to the satellite;
- δP is a matrix of pseudorange observations;
- δU is a navigation error state vector including the receiver's position and clock bias;
- n is a vector containing the pseudo-range measurement noise.

The equation 3, with four visible satellites, takes the following form:

$$\delta U = A^{-1}\,\delta P \qquad (4)$$

If there are five or more satellites, the receiver position is determined using the least squares method. In this case, the equation 3 can be rearranged as:

$$\delta U = (A^T A)^{-1}\,A^T \delta P \qquad (5)$$

where:

- matrix A exemplifies the line of sight vectors that link the receiver to each satellite.

Assuming δU is a zero-mean vector that incloses the errors in the predicted operator state, the term δU estimates the position error. Its covariance can be calculated using the generalized inverse of A [15]:

$$cov(\delta U) = E(\delta U \delta U^T) = \\ E[(A^T A)^{-1}\,A^T\,\delta P\,\delta P^T\,A(A^T A)^{-T}] = \\ (A^T A)\,^{-1} A^T\,dP\,dP^T\,A(A^T A)^{-T} = \\ (A^T A)^{-1}\,A^T\,cov(\delta P) A (A^T A)^{-T} \qquad (6)$$

The term cov(δP) describes the pseudo-range errors, which being uncorrelated, are statistically independent. It follows that the covariance matrix is diagonal. If the variance (σ_n) of the measurement errors for each satellite is the same, the term cov(δP) becomes:

$$cov(\delta P) \,= \sigma_n^2 \qquad (7)$$

By substituting Eq. 7 into Eq. 6

$$E(\delta U \delta U^T) = \sigma_n^2\,(A^T A)^{-1}\,A^T A (A^T A)^{-T} = \sigma_n^2\,(A^T A)^{-T} \qquad (8)$$

Under the assumption that $(A^T A)$ is symmetric, transpose is not necessary. Thus:

$$cov(\delta U) \,=\, \sigma_n^2\,(A^T A)^{-1} \qquad (9)$$

Supposingly $G = (A^T A)^{-1}$, it follows that:

$$cov(\delta U) = \sigma_n^2 \mathrm{G} \qquad (10)$$

The covariance matrix is obtained by expanding the equation:

$$\begin{bmatrix} \sigma_x^2 & cov(x,y) & cov(x,z) & cov(x,b) \\ cov(y,x) & \sigma_y^2 & cov(y,z) & cov(y,b) \\ cov(z,x) & cov(z,y) & \sigma_z^2 & cov(z,b) \\ cov(b,x) & cov(b,y) & cov(b,z) & \sigma_b^2 \end{bmatrix} \qquad (11)$$

$$= \sigma_n^2 \begin{bmatrix} G_{xx} & G_{xy} & G_{xz} & G_{xb} \\ G_{yx} & G_{yy} & G_{yz} & G_{yb} \\ G_{zx} & G_{zy} & G_{zz} & G_{zb} \\ G_{bx} & G_{by} & G_{bz} & G_{bb} \end{bmatrix}$$

The matrix elements quantify the satellite geometry.

B. Dilution of Precision

Dilution of precision is a function of the satellite constellation, more precisely of the geometry between the receiver and satellite. Thus, calculating the DOPs does not require any observations. The DOP values can be predicted using the satellite almanac data or the satellite orbit information. There are several types of DOPs, and each one can be obtained using the diagonal elements of G, independently from the others:

- Horizontal DOP (HDOP): measures the indecision in the horizontal position of the navigation solution:

$$HDOP = \frac{\sqrt{\sigma_x^2+\sigma_y^2}}{\sigma_n} = \sqrt{G_{xx} + G_{yy}} \quad (12)$$

- Vertical DOP (VDOP): corresponds to the uncertainty in the vertical position of the navigation solution;

$$VDOP = \frac{\sigma_z}{\sigma_n} = \sqrt{G_{zz}} \quad (13)$$

- Position DOP (PDOP): represents the uncertainty in the spatial (3D) position of the navigation solution;

$$PDOP = \frac{\sqrt{\sigma_x^2+\sigma_y^2+\sigma_z^2}}{\sigma_n} \quad (14)$$

$$= \sqrt{G_{xx} + G_{yy} + G_{zz}}$$

- Time DOP (TDOP): stands for the uncertainty of the receiver clock;

$$TDOP = \frac{\sigma_b}{\sigma_n} = \sqrt{G_{bb}} \quad (15)$$

- Geometric Dilution of Precision (GDOP): is the combination of all the components that define the impact of geometry on the rapport between the measurement error and the position determination error;

$$GDOP = \frac{\sqrt{\sigma_x^2+\sigma_y^2+\sigma_z^2+\sigma_b^2}}{\sigma_n} = \sqrt{G_{xx} + G_{yy} + G_{zz} + G_{bb}} \quad (16)$$

- Thus DOP terms are interconnected by the following relationship:

$$PDOP^2 = HDOP^2 + VDOP^2 \quad (17)$$

$$GDOP^2 = PDOP^2 + TDOP^2 \quad (18)$$

GDOP is the result of 1/Vol where Vol is the volume of a tetrahedron that is formed linking the receiver position and the four satellites. The best geometry to achieve the highest accuracy for point positioning is when the volume of the tetrahedron is a maximum: this situation needs a minimum value of GDOP [16].

An ideal distribution for four satellites consists of one above the receiver, while the remaining three are separated from each other by 120° in azimuth near the horizon. For this situation, the GDOP value is near to 1.

GDOP is widely used because it is easy to calculate, based on four satellite observations and invariant as the mathematical DOP [17]. Today GDOP is a standard satellite choice because it plays an important role in determining the accuracy of the positioning [18]. All receivers use algorithms that are able to select a subset of at least four satellites with the minimum GDOP.

Even if the minimum number of satellites required to position and estimate clock off-set is four, usually more than four satellites are selected to increase the estimation robustness and to reduce the degradation in the estimate accuracy that is introduced using subsets [19]. This approach is frequently adopted for multi-GNSS constellation.

III. APPLICATION

A. Adopted Approach

The aim of this paper was to calculate a new index, named the Temporal Variability of Geometric Dilution of Precision (TVGDOP), in order to determine the variability in the visibility of satellites in a defined period of time under different scenarios. Simulation tests were performed on a station, located in Naples (Italy). This station is referred to UTM/WGS84 and its coordinates are: φ = 40° 52’ 27’’.05 and λ = 14° 17’ 27”.79 (440263.095, 4525031.994 meters), and it is 81.47 meters above the mean sea level (Fig. 5).

Trimble Planning [20] was used to calculate the GDOP related to date and time. The Ephemeris file containing the orbital parameters concerning each satellite of the considered positioning systems, was downloaded onto the program to evaluate GDOP values.

Our approach is based on three different combinations of positioning systems, with measurements based on:

- GPS;
- GLONASS;
- GPS and GLONASS.

For the above three solutions, satellites availability and GDOP were calculated considering intervals of 15 minutes, for a total of 96 daily samples. The simulations covered a period of 32 days (January 25 – February 25, 2016). To render the simulation more realistic, particular attention was reserved to cut-off angle, the altitude below which a satellite is no visible because obstructed by terrain morphology. Variable values were assumed for cut-off angle: 10°, 15°, 20°, 25°, 30°.

Elevation Cut-off excludes from the calculations all satellites below that height on the horizon. Thus a compensation for the obstruction of those satellites by each specific morphological situation is introduced, i.e. there is assumed to be a clear view of the sky above the elevation cutoff, which means a conic space of satellite visibility. In real surveys, particular obstacles such as neighbouring buildings and trees or mountains (Fig. 6) can mask the GNSS satellites.

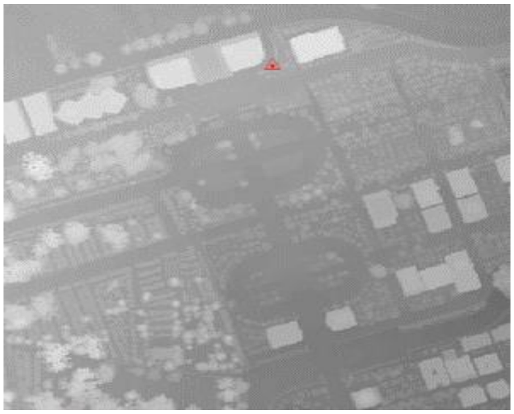


Fig. 5. Station location on the map built from DSM.

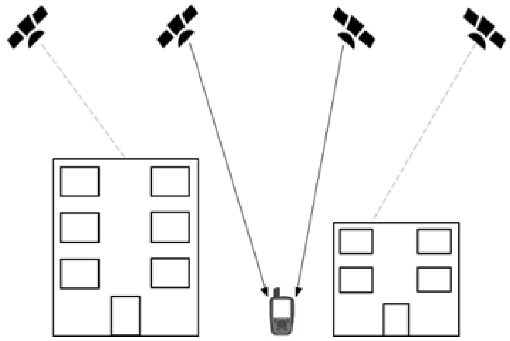


Fig. 6. Example of obstructed view.

Subsequent obstacles that appear along the horizon at the selected location for GNSS measurements are considered. Consequently, another case is investigated, with the same station and the same dates, but with the introduction of an obstruction scenario.

Trimble Planning enables obstacles that appear along the horizon to be drawn using the Obstruction Editor at the location where the GNSS measurements are taken. These include not only the local terrain, but also buildings such as houses or bridges and natural components such as trees. This may be entail reading topographic maps and/or surveying the area. A different and faster approach to define the obstacles caused by the territorial morphology is to use a DSM (Digital Surface Model). A DSM supplies the highest points within a defined grid box [21]. This can be obtained using airborne laser scanning. It is the result of the first echo the laser receives for each laser pulse sent out, and represents the tops of buildings, trees, and other objects, or the ground, if unobstructed [22]. As with DEMs (Digital Elevation Models) [23], a DSM is usually supplied like a grid, i.e. a sample of heights for a number of points that are spaced in regular way. A DTM (Digital Terrain Model) reports the variability of the ground elevation, whereas a DSM depicts the elevation of the top surfaces of buildings, trees, towers and other features elevated above the bare earth [24]. Fig. 7 shows a comparison of a DSM and DTM.

A detailed DSM of the considered area supplied by the local government administration of Naples as a grid interpolation of LIDAR data with cell resolution 1 m was used initially in MicroDEM [25], a free program that works with elevation plots as a function of position. By selecting the position on the map where the GNSS measurements will be taken, the following options are set:

- Max Horizon: How far out it is possible to see, in meters. The Max Horizon will range not beyond than the borders of the DEM.
- Radial precision: Resolution for drawing the radial distances from the considered position as far as the horizon. A shorter distance is more precise, but more time is necessary.
- Angular precision: The misure of the angle between subsequent radials; for this application, 1 degree is set;
- Observer above ground: The vertical distance between the observer and the ground; in this case, the height of the GNSS unit (or its antenna) above the ground.
- Horizon: the characteristics (thickness and color) of the line thar represents the limit of the horizon.

The horizon blocking line is drawn on the map derived from the DSM. The user position is represented by a red square while the horizon blocking line fixes the limit to his field of view (Fig. 8).

The software creates three graphs:

- The first graph (Fig. 9) shows the altitude (the vertical angle) related to the azimuth;
- The second graph (Fig. 10) shows the distance from the user's position to the topographic horizon related to the azimuth;
- The third graph (Fig. 11) illustrates the line of topographic horizon considering a 360 degree wiew.

MicroDEM exports vertical angle values as a function of the azimuth as a text file that can be read by Trimble Planning software. Given with the user's position, the Obstruction Editor plots the text file to detect the occurrence of significant obstacles during the survey. Once the parameters have been set, it is possible to calculate accurate DOPs for the chosen location in relation or not to the obstacles. The flowchart of our approach is reported in Fig. 12.

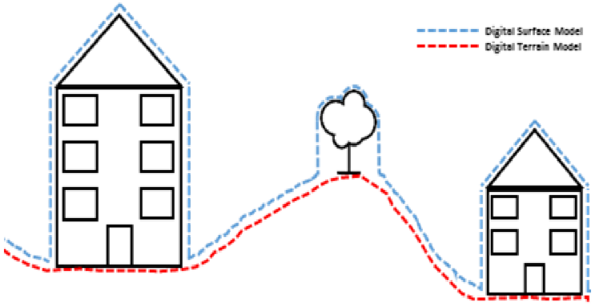


Fig. 7. Examples of a Digital Surface Model and a Digital Terrain Model.

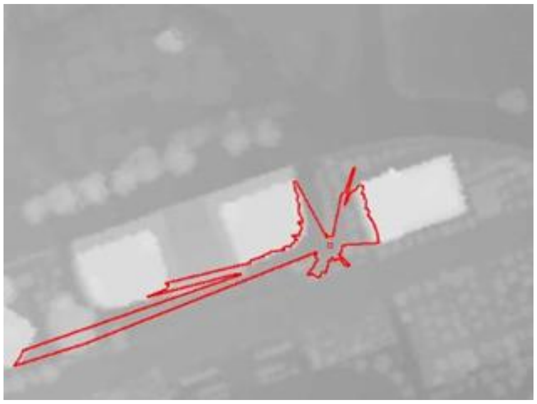


Fig. 8. The Horizon Blocked Area on the Map for the Observer Point.

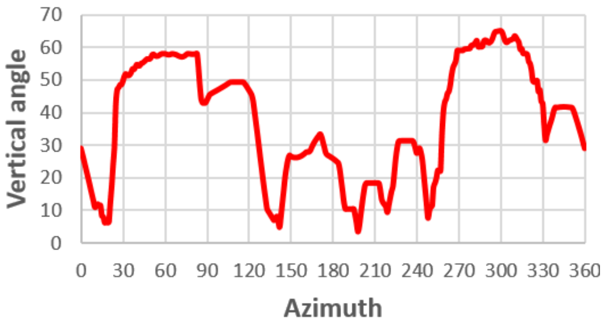


Fig. 9. Representation of Altitude as a Function of the Azimuth.

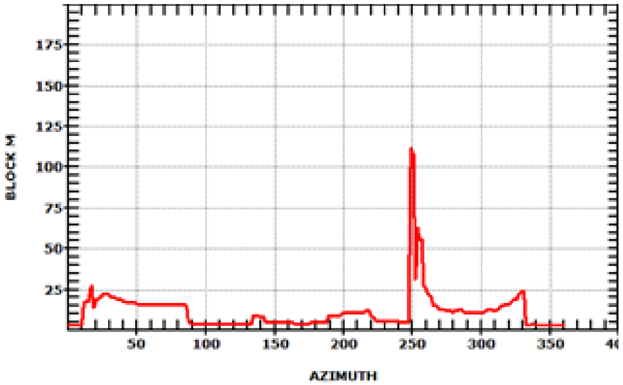


Fig. 10. Representation of Distance from the user's Position to the Topographic Horizon as a Function of the Azimuth.

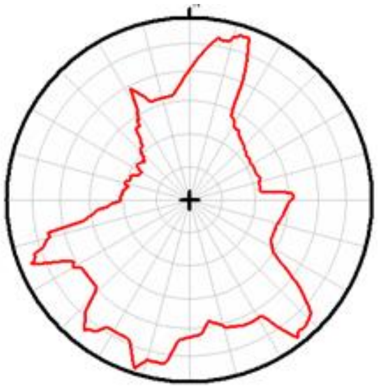


Fig. 11. Topographical Map Shows Area where Blockage Horizon Occurs.

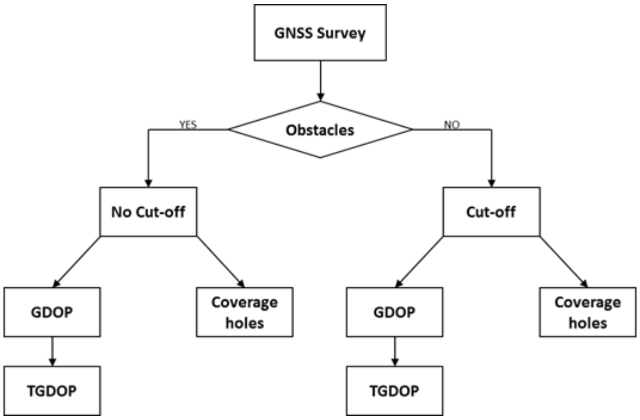


Fig. 12. Flowchart of the Adopted Approach.

In order to determine the GDOP, it is first necessary to decide whether to consider the real obstacle. If they are taken into account, the GDOP calculation method does not consider cut-off angles. If the real obstacles are not contemplated, obstruction scenarios with different elevation cutoff values are studied.

In all cases, the considered period of time is 32 days: for every satellite configuration (GPS, GLONASS, GPS+GLONASS), a minimum collection of available satellites is necessary to achieve reliable GDOP values. More precisely, four satellites for single constellations and five for the GNSS combination are required. For the GNSS, a system time difference parameter is introduced for integrated GPS/GLONASS observation processing [26]. In fact, if the measurements of these different GNSS are combined, the synchronization between the internal GPS and GLONASS receiver clock must be evaluated. Two parameters are contemplated: the unavailability of the minimum number of satellites required for positioning and the temporal variability of GDOP. The unavailability of the minimum number of satellites is determined by the statistical analysis of the daily coverage holes which are the number of times for which less than 4 (for GPS or GLONASS) and 5 (for GPS+GLONASS) satellites are in view during each day. The temporal variability of GDOP is supplied by our proposed index named the Temporal Variability of GDOP:

$$TVGDOP = \sum_{i=1}^{N} \frac{\left(\frac{1}{GDOP_i}\right)}{N} \tag{19}$$

where:

- N is the number of the intervals included in the considered period of time, for example one day (24 hours, 96 intervals).

The introduction of the inverse of GDOP is necessary to contain all values within the interval 0-1. In the worst cases, GDOP is high, thus its inverse is small with only a slight contribution to TVGDOP.

B. Results and Discussion

Tables I and II report the statistics of the coverage holes, the parameter above defined to express (GPS, GLONASS, GPS+GLONASS) satellites visibility, in the absence and presence of obstacles.

The results of Table I highlight how at least four GPS satellites are always visible until the 20 degrees cut-off angle; in fact, coverage holes start to appear after this value although only in small quantities. The GLONASS constellation has the worst performance, presenting coverage anomalies starting at a 15 degrees cut-off. Compared to the GPS, the Russian navigation system has at least twice the number of holes as the corresponding American system. The GPS/GLONASS integration leads to a minimum of five satellites that are continuously in view.

TABLE I. STATISTICS OF THE DAILY COVERAGE HOLES WITHOUT REAL OBSTACLES

Cut-off	**Constellation**	**Coverage holes**		
		Min	**Max**	**Mean**
0° - 10°	*GPS*	0		
	GLN			
	GPS+GLN			
15°	*GPS*	0		
	GLN	0	4	2
	GPS+GLN	0		
20°	*GPS*	0		
	GLN	1	9	4
	GPS+GLN	0		
25°	*GPS*	1	3	2
	GLN	4	12	8
	GPS+ GLN	0		
30°	*GPS*	3	9	6
	GLN	10	23	15
	GPS+ GLN	0		

TABLE II. STATISTICS OF THE DAILY COVERAGE HOLES WITH REAL OBSTACLES

	Min	**Max**	**Mean**
GPS	40	51	44
GLN	64	78	70
GPS+GLN	2	8	5

The statistics in Table II highlight that the obstacle incidence is particularly high with the single constellations. GPS has a minimum daily coverage holes (CHs) value four times higher than its corresponding maximum value without obstacles and with a 30 degrees cut-off angle. GLONASS reveals a worse performance: its mean CHs value (70) is too high considering the total number of daily measurements (96). GPS+GLONASS does not suffer much from this handicap because it has a maximum CH value equal to 8% of the daily observations.

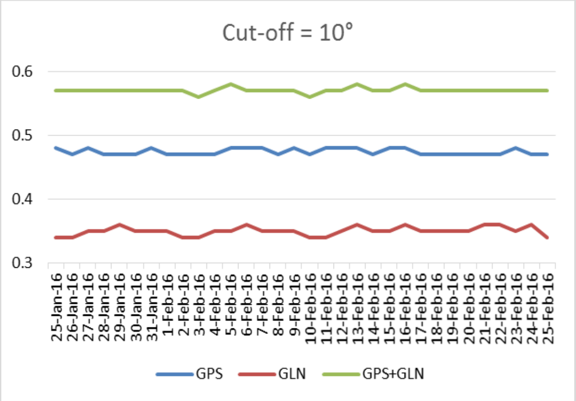


Fig. 13. Daily TVGDOP Values of the Three Constellation Configurations for 10 Degrees Cut-Off.

In order to compare the simulated positioning accuracy of GPS, GLONASS and GPS+GLONASS observations, daily TVGDOP means for the two considered cases (with and without real obstacles), were calculated (Fig. 13 to 18).

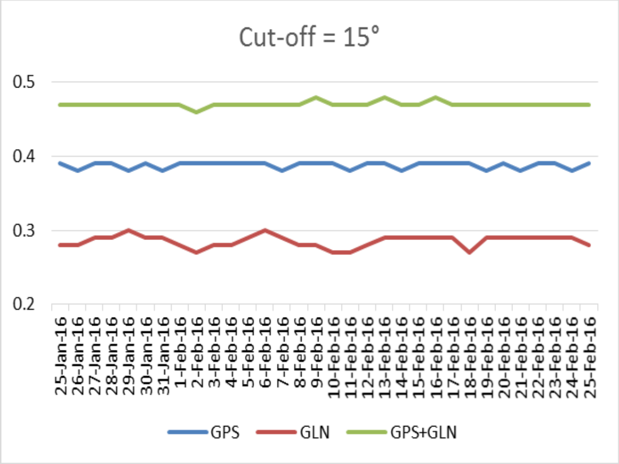


Fig. 14. Daily TVGDOP Values of the Three Constellation Configurations for 15 Degrees Cut-Off.

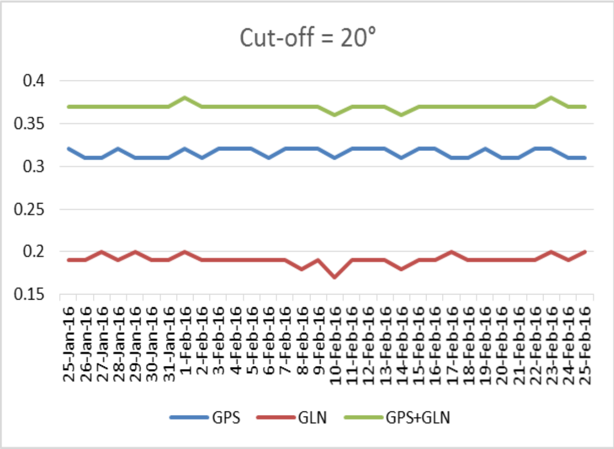


Fig. 15. Daily TVGDOP Values of the Three Constellation Configurations for 20 Degrees Cut-Off.

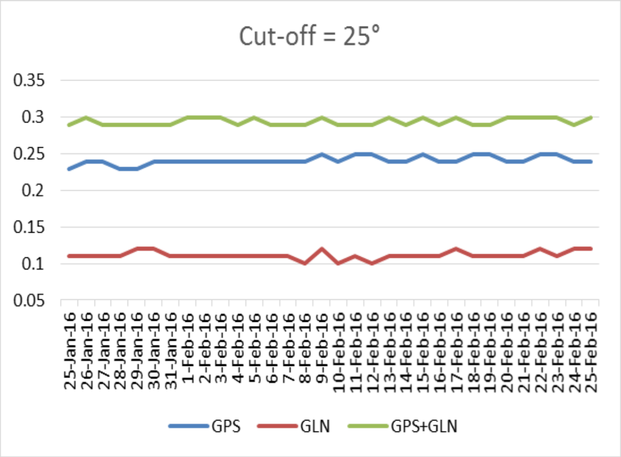


Fig. 16. Daily TVGDOP Values of the Three Constellation Configurations for 25 Degrees Cut-Off.

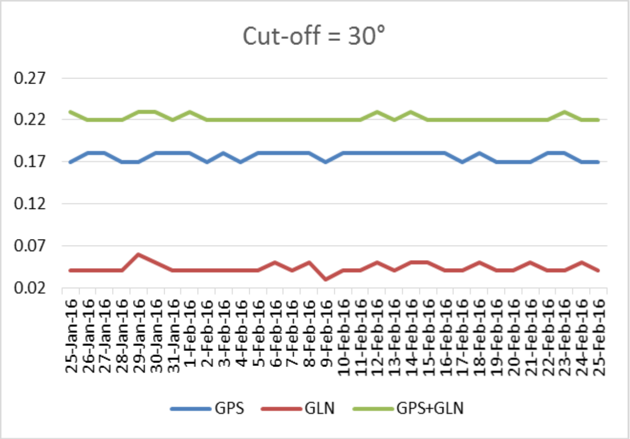


Fig. 17. Daily TVGDOP Values of the Three Constellation Configurations for 30 Degrees Cut-Off.

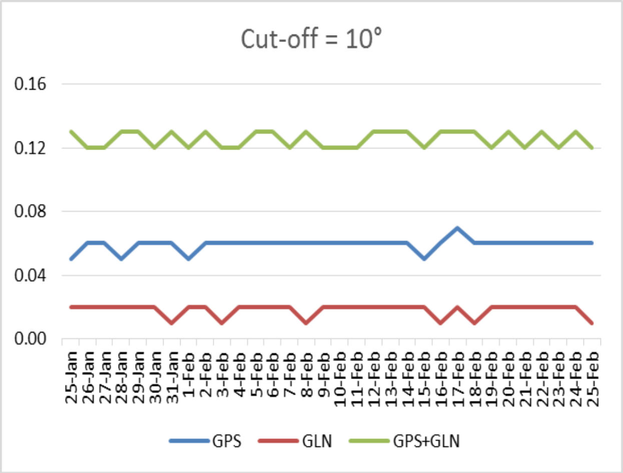


Fig. 18. Daily TVGDOP Values of the Three Constellation Configurations with Real Obstacles.

An increase in a cut-off angle clearly corresponds to a decrease in TVGDOP (Table III).

The index, for the three combinations, is included between the following values:

- 0.18 - 0.47 for GPS;
- 0.04 - 0.35 for GLONASS;
- 0.22 - 0.57 for GPS+GLONASS.

In the presence of obstacles, TVGDOP values decrease: GPS+GLONASS combination only has a value which is approximately 28% of the corresponding case without impediments. However, the impact of the integration is particularly evident because the TVGDOP for GPS+GLONASS presents an increase of 117% compared with GPS.

In all cases, the proposed index well demonstrates that the integration of GPS and GLONASS produces benefits on satellite geometry as it is expected [27] [28] [29].

TABLE III. STATISTICS OF THE DAILY TVGDOP MEAN VALUES

		Cut-off	**Constellation**	**TVGDOP**
Obstacles	**No**	*0°-10°*	*GPS*	0.47
			GLN	0.35
			GPS+ GLN	0.57
		15°	*GPS*	0.39
			GLN	0.29
			GPS+ GLN	0.47
		20°	*GPS*	0.32
			GLN	0.19
			GPS+ GLN	0.37
		25°	*GPS*	0.24
			GLN	0.11
			GPS+ GLN	0.29
		30°	*GPS*	0.18
			GLN	0.04
			GPS+ GLN	0.22
	Yes	*0°*	*GPS*	0.06
			GLN	0.02
			GPS+ GLN	0.13

IV. CONCLUSION

This paper aimed to quantify the impact of the integration of GPS and GLONASS systems on satellite geometry. It presents the latest results obtained within a project carried out at Department of Sciences and Technologies (DiST) of the University of Naples “Parthenope”.

Using the Ephemeris of both constellations for a long period (32 days), the combination of GPS+GLONASS had not only increased the number of visible satellites, but also optimized their space distribution compared to GPS and GLONASS alone. These positive effects were confirmed by the limited or null unavailability of the minimum number of satellites required for positioning and lower values of GDOP, respectively.

To facilitate a comparison between each single system and their integration, we proposed a new index named TVGDOP, which measures the quality of the satellite geometry with reference to a precise time interval (e.g. one day, one week, etc.). The analysis was carried out for different scenarios, in the absence (but with different cut-off angles) and presence (considering DSM) of real obstacles.

The benefits of GPS+GLONASS were more evident in obstructed spaces such as urban environments where a notable increase in TVGDOP was registered compared to a single constellation.

ACKNOWLEDGMENT

The research is supported by University of Naples “Parthenope”.

References

[1] S. Dawoud, GNSS principles and comparison, Potsdam University, 2011.

[2] A. K. Maini, and V. Agrawal, Satellite technology: principles and applications, John Wiley & Sons, 2011.

[3] S.R. Babu, S.I. Dutt, R. Goswami, C.U. Kumari, G.S.B. Rao, and S.S. Rani, "Investigation of GDOP for precise user position computation with all satellites in view and optimum four satellite configuration," Journal of Indian Geophysical Union, vol.13(3), , pp. 139-148, 2009.

[4] C. Cai, and Y. Gao, "Precise point positioning using combined GPS and GLONASS observations," Journal of Global Positioning Systems. vol.6(1), pp. 13-22, 2007.

[5] A. Constantinescu, and R.J. Landry, "GPS/Galileo/GLONASS hybrid satellite constellation simulator-GPS constellation validation analysis," The Institute of Navigation 61st Annual Meeting, pp. 733-737, 2005.

[6] J. Guo, M. Li, X. Li, L. Qu, X. Su, and Q. Zhao, "Precise point positioning with the BeiDou navigation satellite system," Sensors.; 14(1), pp. 927-943, 2014.

[7] Z. Jun, Z. Miaoyan, and Q. Yong, "Satellite selection for multi-constellation," Position, Location and Navigation Symposium, pp. 1053-1059, 2008.

[8] H.A. Karimi, and D. Roongpiboonsopit, "A multi-constellations satellite selection algorithm for integrated global navigation satellite systems," Journal of Intelligent Transportation Systems. vol.13(3), pp. 127-141, 2009.

[9] P. Mattos, "Consumer GPS/GLONASS Accuracy and Availability Trials of a One-Chip Receiver in Obstructed Environments," GPS World, vol.22(12), pp. 32-37, 2011.

[10] M. Fritsche, X. Li, X. Ren, H. Schuh, J. Wickert, and X. Zhang, "Precise positioning with current multi-constellation global navigation satellite systems: GPS, GLONASS, Galileo and BeiDou," Scientific Reports, vol.5(8328), pp. 1-14, 2015.

[11] Government B. Ashman, B., 2020. An Introduction to Global Navigation Satellite Systems, NASA Report GSFC-E-DAA-TN76837, 2019.

[12] L. Zhao, L., VP. áclavovic, and J. Douša, J., "Performance Evaluation of Troposphere Estimated from Galileo-Only Multi-Frequency Observations," Remote Sensing, 12(3), p.373, 2020.

[13] R.B. Langley, "Dilution of Precision," GPS World, vol.10(5), , pp. 52-59, 1999.

[14] J.B.Y. Tsui, Fundamentals of global positioning system receivers, Wiley-Interscience, 2000.

[15] R.G. Brown, P.Y.C Hwang, Introduction to Random Signals and Applied Kalman Filtering, John Wiley and Sons, 1992.

[16] C. Rizos, "Multi-constellation GNSS/RNSS from the perspective of high accuracy users in Australia," Journal of spatial science, vol. 53(2), pp. 29-63, 2008.

[17] A. Krauter, "Role of the Geometry in GPS Positioning," Periodica Polytechnica: Civil Engineering, vol.43(1), pp. 43-53, 1999.

[18] G. Fan, D. Song, C. Xu, and P. Zhang, "An Algorithm of Selecting more than Four Satellites from GNSS," International Conference on Advanced Computer Science and Electronics Information (ICACSEI 2013), Atlantis Press, pp. 134-138, 2013.

[19] J. Zhang, and M. Zhang, "A fast satellite selection algorithm: beyond four satellites," IEEE Journal of Selected Topics in Signal Processing, vol.3(5), pp. 740-747, 2009.

[20] Trimble, Trimble Planning Software, Sunnyvale, USA, 2019.

[21] C. Eberhöfer, M. Hollaus, W. Karel, and W. Wagner, "Accuracy of large-scale canopy heights derived from LiDAR data under operational constraints in a complex alpine environment," ISPRS Journal of Photogrammetry and Remote Sensing, vol.60(5), pp. 323-338, 2006.

[22] R. Behrendt, "Introduction to LiDAR and forestry, part 1: a powerful new 3D tool for resource managers," The forestry source, vol.17(10), pp. 14-15, 2012.

[23] E. Alcaras, C. Parente, and A. Vallario, "A Comparison of different interpolation methods for DEM production", International Journal of Advanced Trends in Computer Science and Engineering. Vol. 6, pp. 1654- 1659, 2019.

[24] H.K. Heidemann, Lidar base specification - Techniques and Methods, U.S. Geological Survey, 2014.

[25] P. Guth, MICRODEM software, Oceanography Department, U.S. Naval Academy, 2017.

[26] K. Fischer, H. Habrich, and P. Neumaier, "GLONASS Data Analysis for IGS," Proceedings of IGS Workshop and Symposium, University of Berne, 2004.

[27] C. Meneghini, and C. Parente, "Advantages of multi GNSS constellation: GDOP analysis for GPS GLONASS and Galileo combinations," International Journal of Engineering and Technology Innovation, 7(1), pp. 01-10, 2017.

[28] C. O'Driscoll, G. Lachapelle, and M. E. Tamazin, "Investigation of the benefits of combined GPS/GLONASS receivers in urban environments," Proceeding on RIN NAV10 Conference on Position, Location, Timing: Everyone, Everything, Everywhere, 2010.

[29] D. Mortari, J. J. Davis, A.Owis, and H. Dwidar, "Reliable Global Navigation System using Flower Constellation," (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No. 2, Feb. 2013.

Enhancing Educational Data Mining based ICT Competency among e-Learning Tutors using Statistical Classifier

Lalbihari Barik[1]*, Ahmad AbdulQadir AlRababah[2], Yasser Difulah Al-Otaibi[3]
Department of Information Systems, Faculty of Computing and Information Technology in Rabigh[1, 3]
King Abdul Aziz University, Kingdom of Saudi Arabia[1, 3]
Department of Computer Science, Faculty of Computing and Information Technology in Rabigh[2]
King Abdul Aziz University, Kingdom of Saudi Arabia[2]

***Abstract*—The implementation of computer-supported collaborative learning has come to play a pivotal role in e-learning platforms. Educational Data Mining (EDM) is a promising area for the exclusive skill development of e-learning tutors, the major concern being investigations over large datasets. The tutors possessing efficient and sufficient soft skills can teach students within less time and with greater productivity. EDM is a regularly used research area that handles the development of methods to explore new ideas in the educational field. Computer-supported collaborative learning in e-learning and competencies on a real-time perspective among teachers are calculated using statistical classifiers. This paper aims to identify a feasible perspective on EDM based ICT competency over e-learning tutors using statistical classifiers. A set of tutors from diverse e-learning centers of various universities is selected for the evaluation purpose. The teachers from the department of mathematics in the universities are selected to attend a professional Qualified Teacher Status numeracy skills test and tutors' online test. The results of online tests are collected and correlated with the Naive Bayes Classifiers algorithms. Naive Bayes Classifiers are used in this paper to find the classification performance results among teachers. Naive Bayes based classification is beneficial for skill identification and improvement among the teachers. Significantly, the data mining classifiers performed well with the large dataset.**

Keywords—Data mining; e-learning tutors; Naive Bayes Classifiers algorithms; ICT; QTS numeracy

I. Introduction

E-learning is a new universal teaching and learning model through the help of electronic resources. A basic teaching model only depends on the help of classrooms and blackboard. The classrooms and blackboard teaching reaches a certain amount of distance only. Other than the class, attendees cannot learn the taught lessons [1]. Computer and internet-based e-learning are introduced to avoid such discrepancies. E-learning is capable of conveying the skill and knowledge to a huge amount of learners at different places. Initially, e-learning is not accepted because they thought that the computer could not lead the human's intelligence in learning [2]. Compare with humans, much information are readily accessible by a single click. E-learning without any intelligent system is not at all possible. Computers are the shapeless information carriers system, so provide any coherent knowledge; some intelligent methods are needed [3].

Data Mining is the best coherent knowledge-based intelligent method. Data Mining is the best field of modern technology in e-learning. Data Mining is the best field of modern technology in e-learning to avoid the inherent difficulty of teaching. Data Mining is a set of data testing methods with the collected data understanding, pre-processing, and reproduction procedure to valuation and implementation [4]. Data mining is an exclusive technique that gives individual concentration and adapts any new modern technique with Information Communication Technology (ICT) as well as the database based data mining technique usually deals with large, heterogeneous, and complex databases. Accordingly, the e-learning database is also well fitted with this description [5].

In large database cases, data mining is the finest procedure, which extorts and identifies the required information and consequent knowledge [6]. These extort and identifications are performed with statistical, mathematical, artificial intelligence, and machine learning techniques [7]. Data Mining is capable of extorting knowledge of the e-learning systems during the testing of information with large datasets. In this work, the most important aim is to discover the ICT skills among the teachers of e-learning centre tutors [8].

Educational Data Mining (EDM) is the best field for the ICT skill development over the e-learning tutors. Computer-supported e-learning and ICT skill among the teachers in real-time are calculated using statistical classifiers. A set of e-learning tutors from various areas is selected for validation purposes. A professional Qualified Teacher Status (QTS) numeracy skills test and tutors' online test based results of mathematical department teachers are selected for validation.

These results are collected and correlated with the Naive Bayes Classifiers algorithms. Naive Bayes Classifiers are used in this paper to discover the classification performance results among the teachers. Naive Bayes based classification is most helpful for skill identification and development amongst the teachers.

*Corresponding Author

II. Literature Survey

The purpose of the data mining method in the e-Learning model is to support the teachers to increase the e-Learning situation. Data mining is the best method to find out valuable information through plenty of techniques such as prediction, classification, rule-based mining, and clustering [9]. DM is the best procedure to analyze the data in a sequence to find out hidden data that are discovering from the previously hidden patterns [10]. DM is a procedure that utilizes the statistical, mathematical methods to extort and recognize the valuable information is as of large databases [11]. Data Mining is used to extract the knowledge of the e-learning method throughout the investigation of the information obtainable in the form of user-generated data. In this work, the main aim is to find out the ICT skill among the e-learning teachers to discover the teachers learning behavior patterns [12].

The data mining technique is fruitfully integrated into the e-learning environment. The advantage of the data mining technique, as well as perception over the e-Learning system, helps to supports the teachers in developing their teaching skills. Data mining is the development of analyzing the input data sequentially to find out the helpful information based on the earlier anonymous data set [13]. Data mining in e-learning intends to afford an advanced picture of the present application of the data mining technique. To propose a practical arrangement of the data mining in e-learning, the collected dataset alone is not possible for the grade or any other criterions predictions. Based on the needed output criteria, the input data are processed with any soft computing technique [14]. The basic existing used techniques for the data processing are the Neural Networks, Genetic Algorithms, Clustering and Visualization Methods, Fuzzy Logic, Intelligent agent-based techniques. In data mining classification methods, the composed data are processed with any one of the following methods, such as clustering-based techniques, classification based techniques, and prediction based techniques. [15].

In the e-learning tutoring process, information and communication technology skills (ICT) are essential. It develops the educational institutions' standards, sustainability-related educational resources, and the selection priority amongst the students in selecting the particular institution [16, 29]. In the beginning, the ICT competency is improved by the Learning Resource Management courses [17]. After that, the teaching system improved to encompass ICT skills with a real-time environment. Nowadays, the world is attractive towards the digitalized, so the teaching and learning methods also enhanced gradually based on digital techniques. In e-learning systems, the ICT skillful teaching methods improve the students learning capacity [18]. ICT in e-learning has many advantages such as easy understanding, easy access to the study materials, time-saving, and book free study [19].

ICT proficiency is necessary to originate their thought of information over some electronic gadgets. Proficiency gives you an idea about the information to locate, evaluate, and use the needed information effectively to teach their subjects [19]. ICT skills are taken into account for the skillful usage of hardware and software, computer system association configuration [20]. Furthermore, the ICT is included with the logical capability with the necessary concept and ability about hardware and software application for efficient utilization of information communication technology [21]. To identify the skills of the tutors are necessary to identify their ranks or ranges. The classification is very much essential to predict the student's ranks. The best classifier is the Neural Network (NN) based statistical classifiers [22].

NN is used to solve real-time statistical problems. NN resolves considerable complicated areas such as business, real estate, education, medicine, and pattern recognition [23]. The reason is that the statistical classifier makes the illustration of the nonlinear function map among the input to output. The statistical classifier has trained through powerful as well as computational efficient technique call backpropagation. The statistical classifier has trained through powerful as well as computational efficient technique call backpropagation [24].

In this work, a decision supported system for ICT skill-based classification on Sequential Minimal Optimization (SMO) in Support Vector Machine (SVM). The SVM based SMO algorithm is designed for testing the ICT skills of the e-tutors in the training process of this model [26, 28]. SVM based SMO algorithm is used for its high accuracy and high speed. The reason is the ease of use and better scaling with the training set with a large dataset. SMO is for evaluating the multipart correlation among input and output of the model [27]. The evaluation results are compared with the SVM based SMO algorithm results. NN results are demonstrated, and intelligent assessments are made by various researchers based on predicting the performance of various departments.

EDM in teaching and learning programs NN has been used to compute the students' performance assessments and monitor them using the Classifiers algorithms [24, 25]. Naive Bayes Classifiers algorithms can calculate the performances of students in the time of admissions to various universities in recent times. The results specify that the Classifier algorithm is potentially enhancing the efficiency of the prediction accuracy of teachers' ICT skills. In this work, a multilayer perceptional recurrent neural network is used with backpropagation learning is utilized. Naive Bayes Classifiers algorithms are used in this study to calculate the ITC among the teachers of the distance education center. The computer science department teachers are recommended to obtain the information and communication technology competency for their education.

III. Methods and Materials

EDM is a talented field of skill development of e-learning tutors, the major concern being investigations over large datasets. The tutors possessing efficient and sufficient soft skills can teach students within less time and with greater productivity. EDM is a regularly used research area that handles the development of methods to explore new ideas in the educational field. Computer-supported collaborative learning in e-learning and competencies on a real-time perspective among teachers are calculated using statistical classifiers. This paper aims to identify a feasible perspective on EDM based ICT competency over e-learning tutors using statistical classifiers. A set of tutors from diverse e-learning centers of various universities is selected for the evaluation purpose. The teachers from the department of mathematics in these universities are selected to attend a professional QTS numeracy skills test and tutors' online test. The results of online tests are collected and correlated with the Naive Bayes Classifiers algorithms. Naive Bayes Classifiers are used in this paper to find the classification performance results among teachers. Naive Bayes based classification is beneficial for skill identification and improvement among the teachers. Significantly, the data mining classifiers performed well with the large dataset. The e-learning data mining process consists of different steps in the general data mining process is shown in Fig. 1.

A. *Educational Data Mining*

EDM is a promising area for discovering the needed predictions from the exclusive type of data sets. EDM is very much helpful for the educational skill-based classifications and the settings of the results based on the classifications. An input of EDM in this work is mining the ICT skills of the e-learning centre tutors. The input data usages of EDM take account of calculating the teacher's performances, as well as lagging skills, which are advised to improve based on the present learning system. In this work, the data mining uses for the investigation and idea of data provides the teaching advice for supports and improves the teacher's skills. This skill improvement brings so many eager learning students to a particular centre.

The first step in EDM is the collection of data. The LMS system is used for collecting, and the interaction information is stored in the database. In this work, we are going to use the e-learning centre teacher's data set. The professional QTS numeracy skills test and tutors' online test information are stored in the database of the DM LMS. In the pre-processing step, the data are arranged in a suitable arrangement to be mined. The database administrator pre-processing tool is used to pre-process the e-learning centre teacher's data set. Then the data are stacked up with an appropriate format.

B. *Processes of Data Mining in e-Learning*

In the e-learning domain, the learners, as well as the teachers, are the soul of the system. The data mining in e-learning is a comfortable area for the special skill development of e-learning center tutors; the major concern is being investigated over large datasets. In e-learning tutors, efficient and sufficient soft skills can teach students within less time and with greater productivity. EDM is a frequently used research field that handles the development of methods to explore new ideas in the teaching-learning area. Computer-supported collaborative learning in e-learning and ICT competencies on a real-time perspective among teachers are calculated using statistical classifiers. This paper aims to recognize a practical perspective on EDM based ICT competency over e-learning tutors using the statistical classifier. E-learning center teachers of various universities are selected for the evaluation purpose. The e-learning center teachers of mathematics in various universities are selected to attend a professional QTS numeracy skills test and tutors' online test.

Data mining has the site records information such as gathering the teacher's profile, real-time accessed information, academic performances of teachers, and the estimation result. Data mining with any online test based result prediction is the main pathway to skill identification and user behaviors. The online test based data mining does much help in the e-learning field to provide real-time instructions for e-learners. Then construct the construct based on the teacher's field of interest. Then build and right to use these fields of interested ICT skills to improve and can implement for the teaching. The results are updated to the data mining server based on the earlier practiced information. Finally, the identified group of teachers of the comparable fields of interest and sending personalized information to the individuals. The proposed e-learning data mining procedure contains three steps are shown in Fig. 2.

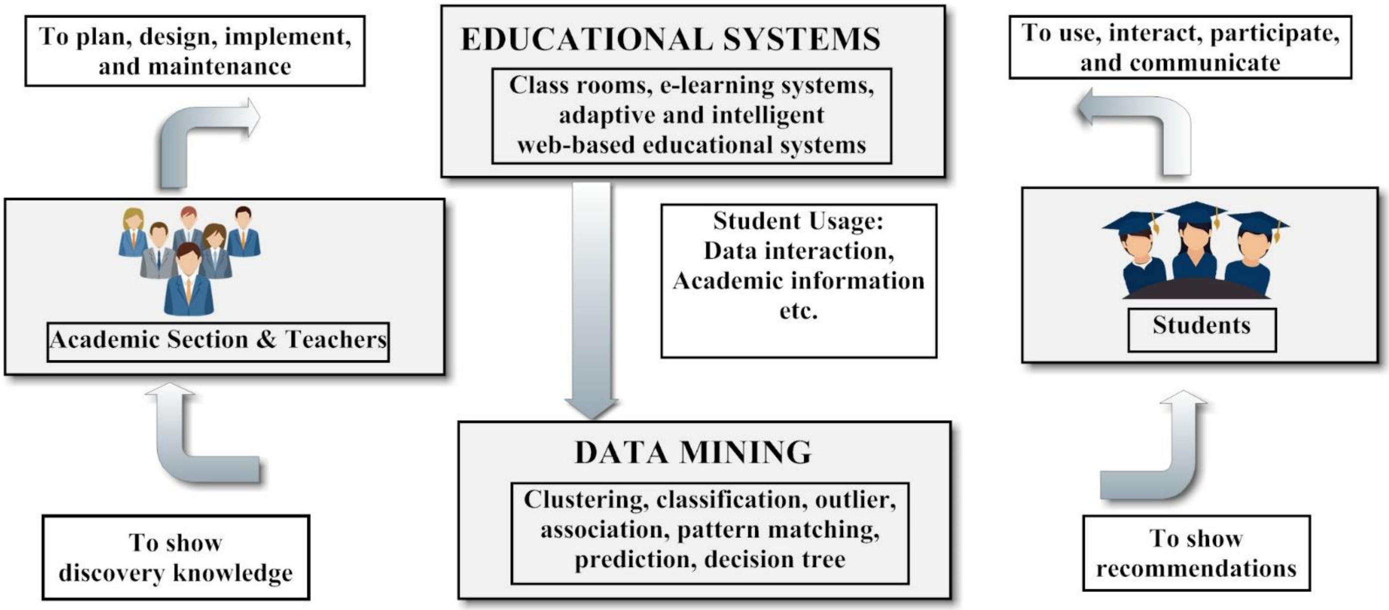

To plan, design, implement, and maintenance
EDUCATIONAL SYSTEMS
Class rooms, e-learning systems, adaptive and intelligent web-based educational systems
To use, interact, participate, and communicate
Student Usage: Data interaction, Academic information etc.
Academic Section & Teachers
Students
DATA MINING
Clustering, classification, outlier, association, pattern matching, prediction, decision tree
To show discovery knowledge
To show recommendations

Fig. 1. Block Diagram of Data Mining in the e-Learning System.

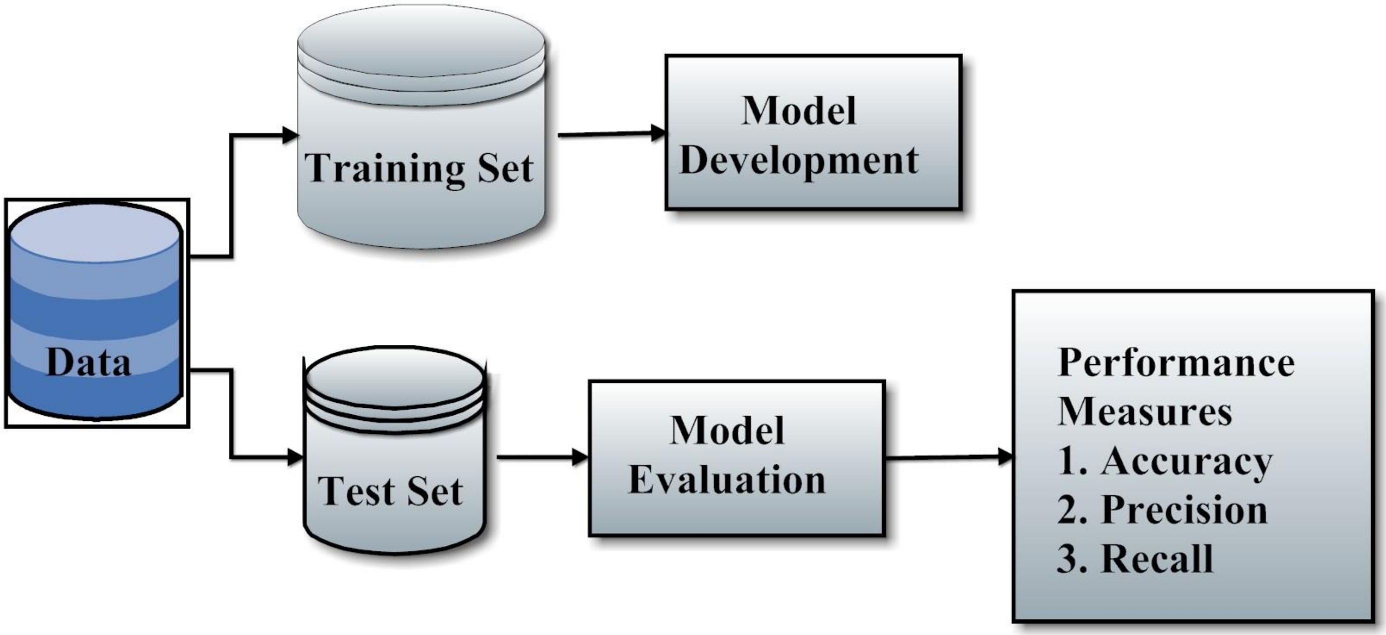


Fig. 2. Data Mining Systems Component.

In this work, data mining with the e-learning development system is iteration based ICT skill of teacher's prediction. Then the minced information has to form some loop of information for classification along with enhanced their teaching skills. Not simply turn the data into relevant knowledge also pass through a filter the mined information for finding the ICT skills among the teachers. The main purpose of data mining in this work is to find the teachers' exact computer skills and their performance together with pedagogical aspects.

C. Naive Bayes Classifier

The Naive Bayes classification algorithm used to construct and implement the data to discover and summarize the information needed for prediction. The data mining algorithm is useful to generate and implement the method that finds out the information and pattern for the ICT skill test. The data mining based classifier is used to find the ICT competency over e-learning centre teachers in this work. The results of the classifiers are used to find the teachers' skills with ICT and skill improvement. The professional QTS numeracy skills test and tutors' online test results are used for deciding the teachers' ICT skill percentage and ICT learning ability. The discovered information are beneficial for the teachers to improve their e-learning system and process. Fig. 3 shows the class-based Naive Bayes Classifier.

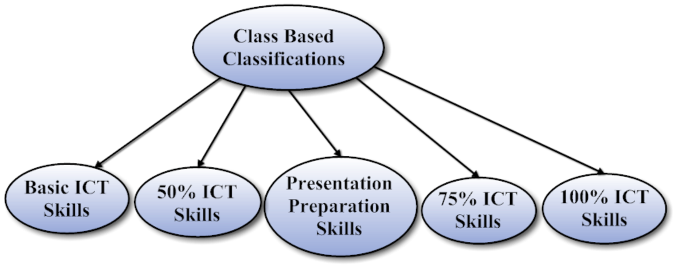


Fig. 3. Naive Bayes Classifier.

The Naive Bayes Classifier belongs to the family of probability classifier with the use of the Bayesian theorem. The reason why it is called 'Naive' because it requires a rigid independence assumption between input variables. Therefore, it is more proper to call Simple Bayes or Independence Bayes. Naive Bayes Classifier is the best method to solve the ICT skill categorization problems. In this work, the preprocessed data are characterized and featured as classes. Based on the results of the skill-based tests, the classifications are performed. The whole data are divided into subclasses such as Basic ICT skills as (0), 50% ICT skills as (1), Presentation Preparation skills (2), 75% ICT skills (3), 100% ICT skills (4). The problem of classifying the class-based probability cases belonging to one category or the other. The goal of the Naive Bayes Classifier is to calculate the conditional probability of the teachers' skill test results, i.e., P (Os |Y1, Y2,…..YN). For every value of P (Os), the possible outcomes of classes are Os. Where 's' is the possible outcomes. Let Y= (Y1, Y2,…,Yn). Using Bayesian theorem, we can get:

$$P(Os|X) = P(Os)P(Os|X)|p(O)\alpha\, P(Os)P(Os|X) = \mathrm{P}\ (\mathrm{Os}\ |\mathrm{Y1}, \mathrm{Y2}, \ldots..\mathrm{YN}) \qquad (1)$$

Then the joint probability values can be written as from can be written in equation (2).

$$\mathrm{P}\ (\mathrm{Os}, \mathrm{Y1}, \mathrm{Y2}, \ldots..\mathrm{YN}) \qquad (2)$$

$$\mathrm{P}\ (\mathrm{Y1}, \mathrm{Y2}, \ldots \mathrm{YN}, \mathrm{Os}) = \mathrm{P}\ (\ \mathrm{Y1}|\mathrm{Y2}, \ldots..\mathrm{YN}, \mathrm{Os}).\ \mathrm{P}\ (\mathrm{Y1}, \mathrm{Y2}, \ldots \mathrm{YN}, \mathrm{Os}) \qquad (3)$$

$$= \mathrm{P}\ (\ \mathrm{Y1}|\mathrm{Y2}, \ldots \mathrm{YN}, \mathrm{Os})$$

$$\mathrm{P}\ (\ \mathrm{Y2}|\mathrm{Y3}, \ldots \mathrm{YN}, \mathrm{Os}).\ \mathrm{P}(\mathrm{Y3}, \mathrm{Y4}, \ldots \mathrm{YN}, \mathrm{Os}) \qquad (4)$$

$$= \ \mathrm{P}\ (\ \mathrm{Y1}|\mathrm{Y2}, \ldots \mathrm{YN}, \mathrm{Os})$$

$$\mathrm{P}\ (\ \mathrm{Y2}|\mathrm{Y3}, \ldots \mathrm{YN}, \mathrm{Os}).\ \mathrm{P}(\mathrm{Y3}|\mathrm{YN}, \mathrm{Os}).\ \mathrm{Os} \qquad (5)$$

Assume that all features Y are mutually independent, we can get:

$$\mathrm{P}\ (\mathrm{Y1}, \mathrm{Y2}, \ldots\ldots \mathrm{YN}, \mathrm{Os}) = \mathrm{P}(\mathrm{Y1}|\mathrm{Os}) \tag{6}$$

Therefore, the formula can be written as:

$$P(Os|X1, X2, \ldots XN)\alpha\, P(OS, \mathrm{Y1}, \mathrm{Y2}, \ldots \mathrm{YN}) \tag{7}$$

$$\mathrm{P}\ (\mathrm{Y1}, \mathrm{Y2}, \ldots \mathrm{YN}, \mathrm{Os}) = \\ \mathrm{P}\ (\ \mathrm{Y1}|\mathrm{OS}). \mathrm{P}\ (\ \mathrm{Y2}|\mathrm{Os}) \ldots \mathrm{P}(\mathrm{YN}|\mathrm{Os}). \mathrm{P}(\mathrm{Os}) \tag{8}$$

$$\mathrm{P}\ (\mathrm{Y1}, \mathrm{Y2}, \ldots\ldots \mathrm{YN}, \mathrm{Os}) = P(Os) \prod_{I=1}^{N} \mathrm{P}(\mathrm{Yi}|\mathrm{Os}) \tag{9}$$

This is the final Naïve Bayes Classifier formula for classification. The class-based predictions are used to calculate parameters and predict the Naive Bayes Classifier. LMS is used to estimate parameters prior to probability and conditional probability.

$$P(Os) = P(\mathrm{Z} = \mathrm{O}s) = \frac{\sum_{x=1}^{M} Q(Zi=OS)}{M} \tag{10}$$

The prior probability equals the number of certain cases of y occurs divided by the total number of records.

$$P(X1 = Bk|z = Os) = \frac{\sum_{x=1}^{M} Q(Y1i=Bk, Zi=Os)}{\sum_{1=1}^{M}(Zi=Os)} \tag{11}$$

The conditional probability of P(Y1=B1|y=C1) equals the number of cases when Y1 equals to a1 and y equals to C1 divided by the number of cases when y equals to C1. Naive Bayes Classifier uses the following formula to make a prediction:

$$\mathrm{Z{=}arg\ max\ P\ (Z{=}Os)} \prod P(Y|Z = Os) \tag{12}$$

IV. Results and Discussion

ICT learning in e-learning and competencies on real-time perspectives among teachers are calculated using the Naive Bayes statistical classifiers. This paper aims to recognize a practicable point of view on EDM based ICT competency over e-learning tutors using statistical classifiers. The tutors from different e-learning centers of various universities are selected for the evaluation purpose. The teachers from the department of mathematics in the universities are chosen to grace with your presence a professional QTS numeracy skills test and tutors' online test. Then the online test are collected and correlated with the Naive Bayes Classifiers algorithms. Naive Bayes Classifiers are used in this work to calculate the classification performance consequences among teachers. Naive Bayes based classification is beneficial for skill classification and development among the teachers. Extensively, the Naive Bayes statistical classifiers performed well with the large dataset.

Table I contains the basic technical aspects particulars of the e-learning center teachers. The teachers of various universities are collected through the internet. Different ICT competencies are investigated and their outcomes are tabulated. In these ICT skill tests, the major three criterions selected are the online test based results, educational qualification, and the collected feedback from the students. From the above table, the ICT skills of the teachers are not only dependent on their educational qualifications. The ICT skills solely depend on the teachers' ICT skills. So the e-learning teachers must have proper ICT skills for e-learning tutoring. Student's feedback on the subject of teacher's presentation percentage are tabulated on the average results. The data of teachers are required to be examined separately to find their ICT skill-based online test. The manual evaluation of the teacher's ICT skill test consumes more time.

The tests are validated by the online test to avoid these kinds of discrepancies. Naive Bayes Classifiers algorithm is necessary to analyze the teacher's data and evaluate their performance.

In the e-learning system, the logical assessments amongst the teachers are performed by the class-based statistical classifier. Then the performances of the teacher's performances over ICT skills are done with the professional QTS numeracy skills test and tutors' online test. The presentation skills and the range of the ICT skills of teachers are classified with the class labels. The ICT skill test data are saved in the DM server. Then the ICT skill test data are arranged based on the performance criteria. Finally, the classification is executed with the five different classes based on the ICT skills over mathematical tests. After the class-based label formations, the results are classified by their performances. The performances of the teachers are selected as excellent, good, medium, normal, and bad performances. Naive Bayes class is a straightforward and high-speed classification algorithm, so it is fit for large datasets. Naive Bayes classification is working with the principle of Bayes theorem of probability for prediction of different classes. Fig. 4 shows the proposed Naive Bayes Classifier.

TABLE. I. Basic Technical Characteristic Particulars

Basic Technical Aspects	**Online Test Results (%)**	**Required Educational Qualification**	**Performance Feedback**
1.a. Basic ICT skills	83%	UG degree, PG degree, Trainees	75%
1.b. Notes downloading ability	85%	PG degree	71%
2.a. Typing skill ability	82%	PG degree	69%
2.b. Media communication ability	84%	Trainees	74%
3.a. Powerpoint preparation ability	87%	UG degree	82%
3.b. Teleconferencing ability	71%	UG degree	80%
4.a. Online test conduction ability	64%	UG degree	68%
4.b. Database management ability	54%	UG degree, PG degree, Trainees	61%
5.a. Computer-Aided animation presentation ability	44%	UG degree, PG degree, Trainees	78%
5.b. Computer-Aided equation making ability	42%	UG degree, PG degree, Trainees	72%

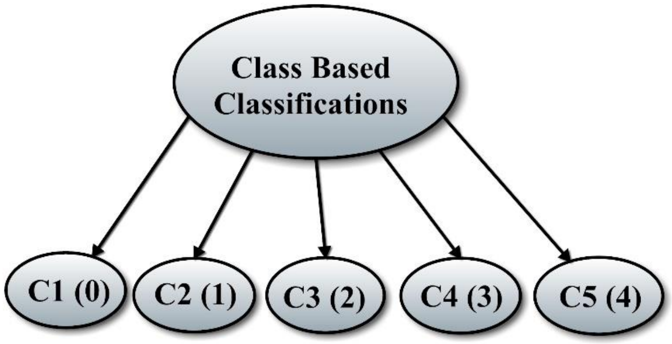


Fig. 4. Proposed Naive Bayes Classifier.

In classification, the initial step is finding the classes which are needed to classify and label them based on performances. Features are characterized with the five classes such as basic ICT skill as C1(0), 50% ICT skills as C2(1), Presentation Preparation skills C3(2), 75% ICT skills C4(3), 100% ICT skills C5(4). This distinctiveness is called features. It is a very useful method to classify teachers. This statistical classification has two input data sets; one is from the online test, and the one is from the professional QTS numeracy skills test. In the testing and training phases, classifiers are used for the evaluation performance. The datasets are divided into test and training set. The validation parameters are validated based on the diverse parameter such as accuracy, precision, and sensitivity.

In the classification part, the ICT skills are classified with the five classifiers that are checked with the depending or not depending based features. If the depending or not depending features are mutually dependent means, it is considered as an independent feature category. This assumption is very important because it reduces computation discrepancy and reduces time consumption. The mutually dependent supposition is named as the class conditional independence. The following equation is the final classifier class-based independence calculation of posterior probability.

$$P(C|F) = \frac{P(F|C)P(C)}{P(F)} \qquad (13)$$

In this work, the classes are based on the ICT skills of the teacher. The probability conditions are calculated based on whether the teachers' skills are in C1(0) or elsewhere. The uses of probability conditions for checking whether the skills are depending only on the 1(a, b). If the case is only on 1(a, b), the class is C1(0). The second case is considered on both skills 1(a, b), 2(a, b) skills are present, means the class is C2(0). The probability P (C|F) is used for the probability finding of the three cases, such as 50% skills, 75% skills, and 100% skill predictions. Which the input skill depends on the second class C2(0) means P (C|F) wants to find the probability of whether the teacher has both the 1(a, b) and 2(a, b) skills. This type of probability that is used to find the priority of the input skill, which depends on the most probable class (F). 'P(C)' is the probability of the skill which is not needed for our prediction or out of our five classes C1(0), C2(1), C3(2)... After the class-based probability predictions, the posterior probability output P (C|F) of our five different labeled classes are predicted. The posterior probability is used for final assumptions of the true class finding of the input skill test data of 'F' given that the class 'C'. The Naive prior probability is used for the assigned five class labels findings without any statistical error. Furthermore, discover the possibility of the class with every characteristic belonging without any error. Then substitute all the predicted class-based probability value in Bayes Formula and calculate posterior probability.

From Table II classes with high probability from the input, value belongs to the high probability classes. In prior and posterior probability calculation, an equation is designed to simplify the probability calculation methods.

Table II is intended to evaluate the posterior probability value. The Posterior probability table encloses the happening of class labels for all characteristics. In the testing and training phases, classifiers are used for the evaluation performance. The datasets are divided into test and training set. Fig. 5 shows that the evaluation parameters are validated based on the diverse parameter such as accuracy, precision, and sensitivity.

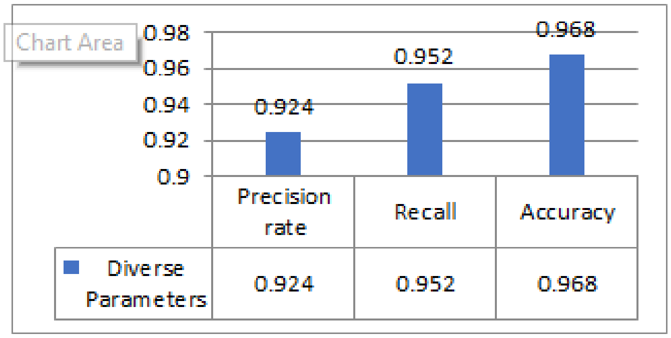


Fig. 5. Evaluation Results Graph.

TABLE. II. POSTERIOR PROBABILITY

Number of class-based/cases	Basic ICT skills as C1(0)	50% ICT skills as C2(1)	Presentation preparation skills C3(2)	75% ICT skills C4(3)	100% ICT skills C5(4)	Posterior probability $P(C\|F) = \frac{P(F\|C)P(C)}{P(F)}$
0.2 (only 1class dependency)	Yes	No	No	No	No	0.00275
0.4 (only 2class dependency)	Yes	Yes	No	No	No	0.0042
0.6 (only 3class dependency)	Yes	Yes	Yes	No	No	0.05920
0.8 (only 4class dependency)	Yes	Yes	Yes	Yes	No	0.7934
1 (all 5classes are dependent)	Yes	Yes	Yes	Yes	Yes	0.910

Precision rate is calculated by the ratio between the correctly labeled positive probable classes among the labeled five classes C1 (0), C2 (1), C3 (2), C4 (3), C5 (4) with the total number of positive classes. Accuracy is the value calculated by the ratio between the fraction of predicted classes among the labeled five classes C1 (0), C2 (1), C3 (2), C4 (3), C5 (4) with the total number of predicted classes. Classification accuracy is the performance evaluator which is used to find the overall performances. Accuracy is the ratio between the total number of accurate class C1 (0), C2 (1), C3 (2), C4 (3), C5 (4) prediction with the total number of overall class predictions done. In the machine learning environment, accuracy, precision, and sensitivity values are essential because these values only showcase our performances over other methods. In this work, the attained accuracy, precision, and sensitivity values are 0.924, 0.954, and 0.968, respectively.

Finally, the results are compared with the SMO of the SVM method. A comparatively naive classifier provides the best performances over SVM because the SVM is a decision-based system only. Initially, the data are tested with the SVM based SMO algorithm for finding the accuracy, precision, and sensitivity. The increased results of the Naïve over SVM are accuracy is over 0.114, precision is increased around 0.09, and sensitivity is over 0.131 values. It shows that our model attained a very standalone performance over SVM. SVM based SMO algorithm is generally used for its high accuracy and high-speed performance. From Fig. 6, we can say that the Naive is performed well over the SVM method. The reason that is Naive can ease of use and better scaling with the training set with large datasets. The best part is the Naive algorithm is executed with less time on large datasets as well as the high accuracy and high-speed performances.

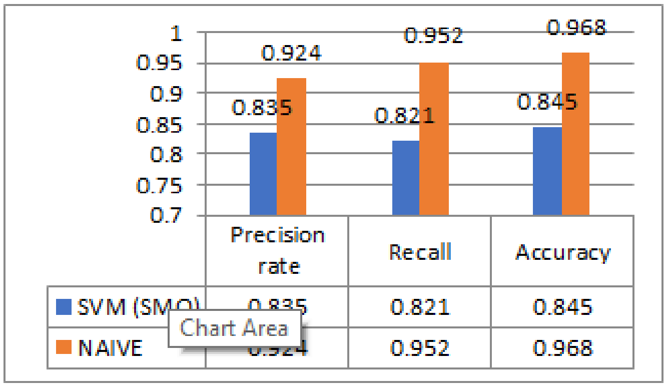


Fig. 6. Comparison with the SMO of the SVM Method.

V. Conclusion

E-learning tutors must have ICT skills over other teachers because they are teaching with the help of computers. The Naive Bayes algorithm is used for the skill prediction because of its simplicity and accuracy over prediction. The most important apprehension is the efficiency over large datasets. Also, this method is well suited to multiple class prediction models. Comparatively, the Naive classifiers perform well over with the logistic regression-based models. The teachers' ICT skills are necessary to teach students within less time and with greater productivity. Computer-supported learning in ICT competencies on a real-time perspective among teachers are calculated with the class-based classifier for efficiency improvement. In this work, a feasible perspective of the posterior probability-based statistical classifier is used for error-free prediction. The teachers from the department of mathematics are not fluent in computer skills. To improve them by knowing them with their result range is very easy to give them training regarding soft skills. Naive Bayes based classification is an appropriate skill identification and improvement among the teacher with large datasets.

Further, the dependent and independent features are taken into account. In practice, individual dependent and independent model alone model statistical predictions are not possible because the prediction depends only on both the features only. The necessities of predictors are must be independent features. According to the real-time case, the predictors are only dependent only on the performance of the classifier. Additionally, classification accuracy can also improve through the use of noise elimination techniques to get rid of an outlier in data sets.

Acknowledgment

The author thanks King Abdulaziz University for this work.

References

[1] Anduela Lile, "Analyzing E-Learning Systems Using Educational Data Mining Techniques," Mediterr. J. Soc. Sci., vol. 2, no. 3, pp. 403-419, 2011. doi: http://dx.doi.org/10.5901/mjss.2011.v2n3p403.

[2] F. Castro, A. Vellido, À. Nebot, and F. Mugica, "Applying Data Mining Techniques to e-Learning Problems," Studies in Computational Intelligence (SCI) vol. 62, no. 221, pp. 183–221, 2007.

[3] Romero, Cristobal, and Sebastian Ventura, eds. "Data mining in e-learning," WIT Press, Vol. 4, 2006.

[4] L. Jiang, H. Zhang, and Z. Cai, "A Novel Bayes Model : Hidden Naive Bayes," IEEE Transaction on Knowledge and Data Engineering, vol. 21, no. 10, pp. 1361–1371, 2009. doi:http://dx.doi.org/10.1109/TKDE. 2008.234.

[5] S. Taheri and M. Mammadov, "Learning the naive bayes classifier with optimization models," Int. J. Appl. Math. Comput. Sci., vol. 23, no. 4, pp. 787–795, 2013. doi:http://dx.doi.org/10.2478/amcs-2013-0059.

[6] S. Lawe and R. Wang, "Optimization of Traffic Signals Using Deep," AI 2016 Adv. Artif. Intell. AI 2016. Lect. Notes Comput. Sci., vol. 9992, pp. 403–415, 2016. doi:http://dx.doi.org/10.1007/978-3-319-50127-7 27.

[7] W. Zhang and F. Gao, "An improvement to naive bayes for text classification," Procedia Eng., vol. 15, pp. 2160–2164, 2011. doi:http://dx.doi.org/doi:10.1016/j.proeng.2011.08.404.

[8] S. Banga, S. Mongia, S. Dhotre, and I. Introduction, "Regression And Augmentation Analytics on Earth ' s Surface Temperature," vol. 5, no. 3, pp. 17–19, 2017.

[9] X. Wu et al.," Top 10 algorithms in data mining," Springer-Verlag London, vol. 14, no. 1. 2008. doi:http://dx.doi.org/doi:10.1007/s10115-007-0114-2.

[10] A. Choi, N. Tavabi, and A. Darwiche, "Structured features in naive bayes classification," 30th AAAI Conf. Artif. Intell. AAAI, 2016, pp. 3233–3240, 2016.

[11] Zhang H., "The Optimality of Naive Bayes," 2004. American Association for Artificial Intelligence (www. ,. org). 2004.

[12] T. Calders and S. Verwer, "Three naive Bayes approaches for discrimination-free classification," Data Min. Knowl. Discov., vol. 21, no. 2, pp. 277–292, 2010. doi:http://dx.doi.org/doi:10.1007/s10618-010-0190-x.

[13] Laney D., "3D data management: Controlling data volume, velocity and variety," META group research note. 2001 Feb 6; 6(70):1, 2001.

[14] R. Y. M. Li and H. C. Y. Li, “Have housing prices gone with the smelly wind? Big data analysis on landfill in Hong Kong,” Sustain., vol. 10, no. 2, pp. 1–19, 2018. doi:http://dx.doi.org/doi:10.3390/su10020341.

[15] Boyd, Danah and Crawford, Kate, “Six Provocations for Big Data,” A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 2011. Available at SSRN: https://ssrn.com/abstract=1926431 or http://dx.doi.org/10.2139/ssrn.1926431.

[16] K. Swan and L. F. Shih, “on the Nature and Development of Social Presence in Online Course Discussions,” Online Learn., vol. 9, no. 3, pp. 115–136, 2019.

[17] Segaran, Toby, and Jeff Hammerbacher, “Beautiful data: the stories behind elegant data solutions,” O'Reilly Media, Inc., p. 257, 2009.

[18] S. O. Material, S. Web, H. Press, N. York, and A. Nw, “The World ’ s Technological Capacity,” vol. 60, no. 2011, pp. 60–66, 2014. doi: http://dx.doi.org/10.1126/science.1200970. PMID 21310967

[19] P. Larrañaga, H. Karshenas, C. Bielza, and R. Santana, “A review on evolutionary algorithms in Bayesian network learning and inference tasks,” Inf. Sci. (NY)., vol. 233, pp. 109–125, 2013. doi:http://dx.doi.org/10.1016/j.ins.2012.12.051.

[20] M. E. Maron, “Automatic Indexing: An Experimental Inquiry,” J. ACM, vol. 8, no. 3, pp. 404–417, 1961. doi: http://dx.doi.org/10.1145 /321075.321084.

[21] Rennie, J.; Shih, L.; Teevan, J.; Karger, D., “Tackling the poor assumptions of Naive Bayes classifiers,” Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

[22] Rish, Irina. "An empirical study of the naive Bayes classifier." IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. 2001.

[23] Russell SJ, Stuart J. Norvig. Artificial Intelligence: A Modern Approach. 2003:111-4.

[24] R. Caruana and A. Niculescu-Mizil, “An empirical comparison of supervised learning algorithms,” ACM Int. Conf. Proceeding Ser., vol. 148, pp. 161–168, 2006.

[25] Keerthi S. S., Shevade S. K., Bhattacharyya C., and Murthy K. R. K., “Improvements to Platt’s SMO Algorithm for SVM Classifier Design,” Neural Computation, 13: 637-649, 2001.

[26] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy, “Improvements to the SMO algorithm for SVM regression,” IEEE Trans. Neural Networks, vol. 11, no. 5, pp. 1188–1193, 2000.

[27] N. Matić, I. Guyon, L. Bottou, J. Denker, and V. Vapnik, “Computer aided cleaning of large databases for character recognition,” Proc. - Int. Conf. Pattern Recognit., vol. 2, pp. 330–333, 1992.

[28] Ali Z, Shahzad SK, and Shahzad W. Performance analysis of support vector machine based classifiers. International Journal of Advanced and Applied Sciences, 5(9): 33-38, 2018. doi:http://dx.doi.org/doi:0.21833 /ijaas.2018.09.007.

[29] C. Jensen, M. Kotaish, A. Chopra, K. A. Jacob, T. I. Widekar, and R. Alam, “Piloting a Methodology for Sustainability Education: Project Examples and Exploratory Action Research Highlights,” Emerg. Sci. J., vol. 3, no. 5, pp. 312–326, 2019. doi:http://dx.doi.org/10.28991/esj-2019-01194.

Histogram Equalization based Enhancement and MR Brain Image Skull Stripping using Mathematical Morphology

Zahid Ullah[1], Prof. Su-Hyun Lee[2]*, Prof. Donghyeok An[3]
Department of Computer Engineering
Changwon National University
Changwon, South Korea

***Abstract*—In brain image processing applications the skull stripping is an essential part to explore. In numerous medical image applications the skull stripping stage act as a pre-processing step as due to this stage the accuracy of diagnosis increases in the manifold. The MR image skull stripping stage removes the non-brain tissues from the brain part such as dura, skull, and scalp. Nowadays MRI is an emerging method for brain imaging. However, the existence of the skull region in the MR brain image and the low contrast are the two main drawbacks of magnetic resonance imaging. Therefore, we have proposed a method for contrast enhancement of brain MRI using histogram equalization techniques. While morphological image processing technique is used for skull stripping from MR brain image. We have implemented our proposed methodology in the MATLAB R2015a platform. Peak signal to noise ratio, Signal to noise ratio, Mean absolute error, Root mean square error has been used to evaluate the results of our presented method. The experimental results illustrate that our proposed method effectively enhance the image and remove the skull from the brain MRI.**

Keywords—Contrast enhancement; skull stripping; magnetic resonance imaging; mathematical morphology

I. Introduction

Magnetic resonance imaging (MRI) is a non-invasive and an important imaging technique. MRI offers a high distinction of various soft tissues. Different applications of brain MRI, such as brain tissue segmentation, pathology detection, and multi-modal brain image registration need to extract the brain region as a preliminary step.

Medical brain imaging application is extensively used to detect different brain abnormalities like for instance paralysis, stroke, breathing difficulties and brain tumor. Skull stripping is an important pre-processing step in brain imaging since the last decade or so [1]. Segmentation is an important tool to study and diagnose various diseases such as autism [2], Alzheimer disease [3-5], and epilepsy [6] and it also required the recognition of anatomical structures, and brain tissue classification. The automatic skull stripping of a brain MRI is a challenging task due to low contrast, unclear brain boundaries, and pixels similarity. Using MRI datasets with a pathological disorder, the entire brain extraction becomes more challenging and problematic [37]. Nowadays, deep learning techniques such as convolution neural network-based algorithms, are mostly used to overcome medical imaging problems [38]. CNN based algorithm [39] employed known labeled data to learn the mathematical description required for the region or object detection, segmentation and classification.

The skull stripping removes skull, scalp, dura and skin/muscle from MR images for keeping only cerebral tissues. In MR image several brain diseases look similar specifically that disease which has an impact on cerebral atrophy. Those sensitive details cannot be differentiated by human naked eyes. Therefore, the enhancement of an image is necessary to accurately identify those details. The state-of-the-art skull-stripping can be divided into four different classes, such as deformable surface model [7-10], thresholding with morphology [11, 12], region-based [13-15] and hybrid approaches [16-24].

The deformable surface model initially expresses the surface model and then repeatedly deforms the surface until it found the optimal solution. Brain extraction tool proposed by Smith [10], this method does not require any pre-processing/pre-registration before implementation. The region-based segmentation methodology takes the brain part as a single connected region. The region-based segmentation merges the same region into one larger region. This method comprised of watershed techniques [14], region growing [13,15], etc. The hybrid technique integrated different existing methods to enhance performance.

The scan of the brain MRI consists of Axial, Coronal, and Sagittal view as shown in Fig. 1. We can obtain the Axial view by dividing the brain by a horizontal/ lateral plane. The Axial view divides the brain into two parts such as inferior and superior parts. However, the brain is divided into a Coronal view into ventral and dorsal parts by vertical/frontal plane. While, in the Sagittal view, the brain MR image is obtained by dividing the brain into right and left parts of the longitudinal/median plane. For different brain MR image applications, skull stripping is an important step like for instance brain strokes, brain tumor segmentation, Dyslexia, Epilepsy, and brain tissue segmentation. Also, the skull stripping method is used to remove the non-brain tissues such as skull, eyes, dura, scalp, etc. from MR brain images.

*Corresponding Author
E-mail: sleepl@changwon.ac.kr

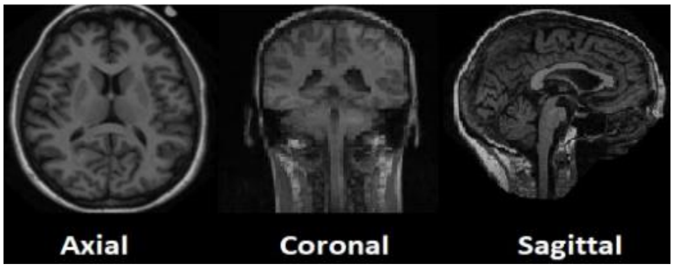


Fig. 1. Three Different views of the Brain.

In addition, there are numerous brain extraction algorithms, but the results of these algorithms on all brain MRI data sets are not satisfactory. Therefore, a robust and fully automated brain extraction algorithm is required which extracts the brain part accurately from a brain MRI database. MR brain volume exhibits numerous imaging artifacts, due to the limitations of material heterogeneity and spatial resolution of imaging modality, such as blurring, noise, partial volume effect, inhomogeneity and so on. Due to this imaging artifact, brain extraction becomes more difficult.

Considering all these limitations, we have proposed an accurate and robust skull stripping algorithm in this paper. We have used morphological based methods to remove the skull from the image while the histogram equalization based techniques have been used to enhance the brain MR image. As the magnetic resonance imaging modality generates the low contrast image and in this low contrast image it is very difficult for a doctor or radiologist to diagnose a disease. If the contrast of an image is high so the detail information can be easily analyzed. Therefore, we have enhanced every MR image as a pre-processing step using different histogram equalization techniques. In the human head, the skull is the hardest part and the skull act as a protector for the brain. However, diagnosing different brain diseases, the skull is a redundant part and it must be removed from the image. Therefore we have proposed an efficient methodology for skull stripping using mathematical morphology technique. In literature [25-28], several methods such as semi and fully automated methods are presented for skull removal.

II. Brain MR Image

Magnetic resonance imaging is the most effective imaging modality to study the brains because the MRI has the capability to image the brain structures both interior and exterior with a high spatial resolution image of anatomical details, therefore, minute changes can be read or detected in these structures. The magnetic resonance imaging can generate images in any direction from side to side, top to bottom, or front to back. That’s why the three-dimensional brain MR images are getting popular day by day in medical applications and also being used for research-related treatment, diagnosis, image-guided surgeries, and surgical planning.

The MR brain images are having three different types of images, T1-weighted, T2-weighted, and PD-weighted, where each type of MR brain image is focused on various contrast characteristics of the brain tissues [33]. The magnetic resonance brain imaging has got so many advantages as compared to other imaging modalities. The brain MR images are clearer and showing more detail related to other existing imaging modalities. MR imaging has the tendency to image the brain in any plane without moving the patient physically whereas computer tomography is limited to only one plane, which is an axial plane [34-35]. Due to these features, the MRI is an invaluable tool to evaluate or diagnose different brain diseases.

The MR brain imaging has been extensively used to diagnose different brain diseases, like Alzheimer disease, arteriogram, blood clots, brain tumor, Huntington's disease, hypopituitarism, stroke, multiple sclerosis, optic glioma, petit mal seizure, partial seizure, subdural hematoma, Cushing disease, etc. [36]. There are so many brain diseases, some of them are shown in Fig. 2.

The structure of the rest of the review paper is as follows: Section 3 presents the literature review, Section 4 presents the proposed model, Section 5 presents results and discussion, Section 6, presents Conclusion and Future Work.

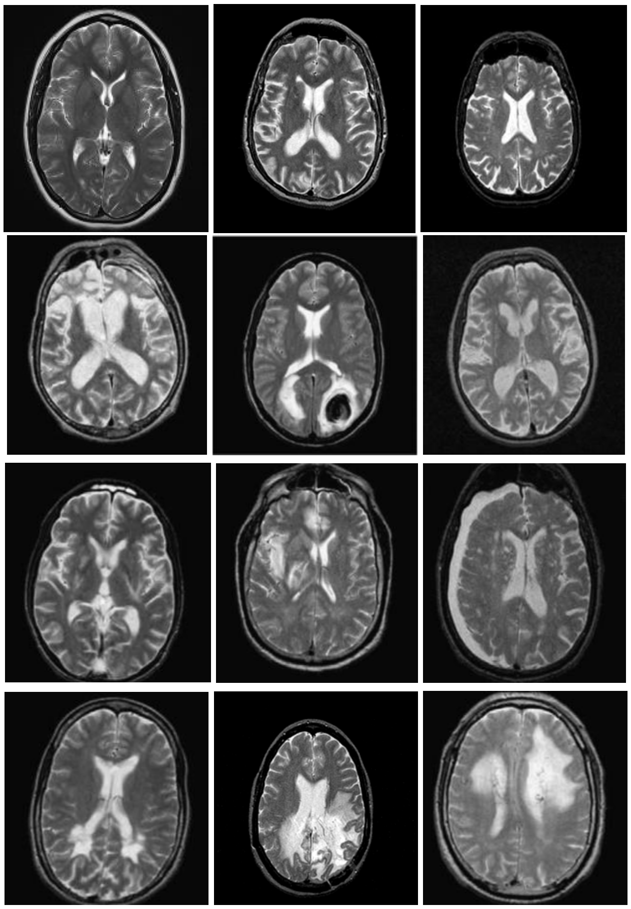


Fig. 2. Different brain MR images samples: a) Normal, b) Alzheimer's disease, c) Alzheimer's disease plus visual agnosia, d) Pick's disease, e) sarcoma, f) Huntington's disease, g) cerebral toxoplasmosis, h) herpes encephalitis, i) chronic subdural hematoma, j) multiple sclerosis, k) glioma, and l) meningioma.

III. Literature Review

Iglesias et al. [29] presented ROBEX (robust learning-based brain extraction tool). The generative and discriminative models are combined after standardizing signal intensities and bias filed correction is implemented. BEaST (Brain Extraction based segmentation) is another contemporary method [30]. Spatial and intensity normalization of the data is important in this method. The current methods are effective for some specific datasets but unfortunately not appropriate for others. Mathematical morphology proposed by Gonzales and woods [31] is an effective methodology for extracting skeleton, convex hull and other boundaries. For brain segmentation and analysis mathematical morphology has been used by different researchers [7,11]. They have used the morphological opening for brain tissue separation from the surrounding tissues while morphological closing and dilation have been employed to fill the holes in the image. As for further processing of an image, a binary form image is required for morphological operation. From the gray level image, the threshold creates a binary image by converting all the pixels values to zero which are below the threshold and those pixels which are above the threshold value are considered as one. However, the selection of a robust threshold value is a challenging task. In [16-24], hybrid approaches have been used for extracting the initial brain region, morphology-based method, in these cases, they have used the intensity thresholding. Further, the final binary brain mask is generated for various morphological tasks. The selection of accurate threshold value in these approaches is very challenging to find the region of interest. In reference [40], a survey can be seen on all the existing conventional skull-stripping methods. The state-of-art work can be seen in reference [41-47].

IV. Proposed Methodology

MRI is the most effective imaging modality to study the brain tissues thoroughly as the MRI has the capability to capture the image structures both internally and externally with a high spatial resolution of anatomical details, therefore, minute changes can be read or detected in these structures. There are so many applications of brain imaging in the medical science field [7]. For these applications, the MR images are commonly used. In this paper, we have presented a robust algorithm for contrast enhancement and skull removal.

V. Contrast Enhancement

The MR imaging modality usually generates low-quality images and extracting information from a low-quality image is not an easy task. Therefore, in the first stage of our proposed methodology, we have presented an efficient technique for MRI contrast enhancement based on histogram equalization techniques such as:

A. *Median Filter*

The median filter is used in the pre-processing stage to the MR brain image for the removal of salt and pepper noise. As the MR image consists of salt and pepper and rician noises. The median filter removes the noises from MR images effectively while preserves the edges of the image efficiently. The median filter is a non-linear filter and this filter proceeds in such a way where it considers every pixel by the median value of the neighboring pixel. We have used a 3 x 3 window size for image filtering, as this window size is a suitable window size to filter an image.

B. *Histogram Equalization*

The HE can be represented as the mapping or transformation of every pixel of the input image into corresponding pixels of the processed output image [31]. The function of histogram equalization is to adjust the image intensities to improve the image contrast. The equation of histogram equalization is as follows:

$$P_n = \frac{no.of\ pixels\ with\ intensity\ n_k}{total\ no.of\ pixel\ n} \qquad (1)$$

The range of the MR gray level image is [0… L-1].

C. *Adaptive Histogram Equalization*

In the fourth stage of the proposed methodology, we have used AHE, as this technique is effective for medical images to enhance the contrast of the image. Adaptive HE does not apply transformation or mapping on the overall image, but it performs separately on the sub-image and then combine the image in a proper way.

Pseudo-code of our proposed methodology is given below:

Algorithm: Brain MRI Enhancement and Skull Stripping

```
Input: brain MR image
Parameter: N is the total number of images
Step 1 (Median filter, HE, AHE, and CLAHE
For I = 1 : N
Read the images and implement the above techniques //Enhanced Image
End
Step 2 Skull Stripping using mathematical morphology
For I = N
Read the enhanced brain MRI
Binarized the image using Otsu thresholding
Extract the largest connect component from the binary image
End
Output: Skull Stripped Image
```

D. *Contrast Limited Adaptive Histogram Equalization*

CLAHE is an extension of the adaptive HE technique [36]. CLAHE and AHE are specifically used to curb the over-enhancement problem of HE. CLAHE is employed to control the noise problem which is existed in traditional histogram equalization. In the MRI image, CLAHE works on the small regions which are known as tiles and it also calculates different histograms, and then compares each histogram to a specific part of the image and furthermore, it is utilized to reorganize the contrast estimation or brightness of the image. CLAHE provides more details as compare to standard histogram equalization as CLAHE improves the contrast of the image effectively but CLAHE still has the inclination to amplify unwanted pixels that have to be improved in the future work. The enhanced result of the gray level *l* is computed by employing the below equation:

$$Yl = T(l) = \frac{R-1}{N}\sum_{k=o}^{l} H(k) \tag{2}$$

Where $T(l)$ illustrates the mapping function and plots the different levels *l* of the input picture into y_l .

VI. Skull Stripping

In MR brain imaging application the removal of the skull is a major stage and separation of non-cerebral tissues from cerebral tissues is known as skull stripping. In the skull removal process, the key problem is the separation of intracranial and non-cerebral tissues due to their intensities similarity. So we have presented an efficient methodology to overcome this issue by employing a mathematical morphology-based method as shown in Fig. 3.

A. *Otsu Thresholding*

The Otsu algorithm uses the zeroth and the first order cumulative moment of the gray level histogram. This algorithm is one of the simplest algorithms and is shown as follows:

$$P_i = \frac{n_i}{N}, P_i \geq 0, \sum_{i=1}^{L} P_i = 1 \tag{3}$$

B. *Mathematical Morphology and Hole Filling*

The morphological operations are implemented on a binary image such as erosion, dilation, and region filling to separate redundant areas. The binary image is convolved with a structuring element to generate the skull removal picture. As the structure of the brain is like an oval shape, therefore we consider a disk-shaped structuring element in the process of convolution as shown in Fig. 4.

We have used erosion to remove the pixel's which are residing on the boundaries of brain MR image and is also used for the elimination of non-brain regions such as meninges and skull. In reference [31] explains the erosion of a binary image as follows, *A* employs structuring element while *B* can be represented as follows:

$$\mathrm{A}\Theta\mathrm{B} = \{\mathrm{Z} \,|\, (\mathrm{B})_z \subseteq \mathrm{A}\} \tag{4}$$

The above equation can be explained as [31], erosion of A by B is the set of all points z such that *B*, translated by z, is contained in *A.* While dilation can be defined as,

$$\mathrm{A} \oplus \mathrm{B} = \{\mathrm{Z} \text{ such that}(\hat{B})_z \cap \mathrm{A} \neq \varnothing\} \tag{5}$$

The morphological dilation is employed in the image to unite entire intracranial tissues in the picture and can be explained as, dilation of A by B is the set of all displacements, z, such that $\hat{B}$ and *A* overlap by at least one element.

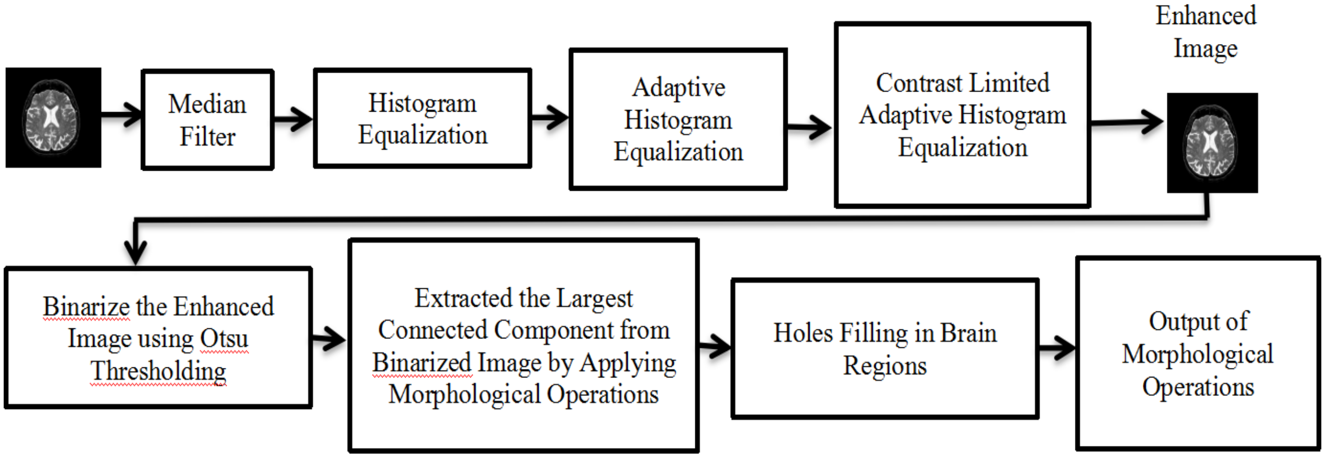
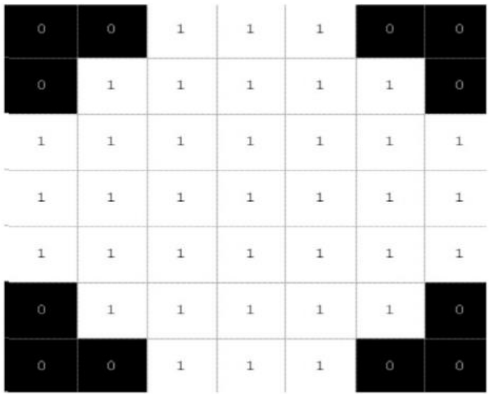

Enhanced Image
Median Filter
Histogram Equalization
Adaptive Histogram Equalization
Contrast Limited Adaptive Histogram Equalization
Binarize the Enhanced Image using Otsu Thresholding
Extracted the Largest Connected Component from Binarized Image by Applying Morphological Operations
Holes Filling in Brain Regions
Output of Morphological Operations

Fig. 3. Proposed Methodology for Brain Image enhancement and Skull Stripping.



0	0	1	1	1	0	0
0	1	1	1	1	1	0
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
0	1	1	1	1	1	0
0	0	1	1	1	0	0

Fig. 4. Morphological Erosion and Dilation Structuring Element.

VII. Results and Discussion

The simulation has been carried out using MATLAB 2015a to validate the proposed scheme on a personal computer with 3.30 GHz Core-i5 processor and 4 GB RAM, running under Windows 10 operating system. Different types of medical image databases are being used for image segmentation. In this study, the database of brain MR images is collected from Harvard Medical School website, and the URL is http://www.med.harvard.edu/aanlib/home.html [32].

The magnetic resonance imaging modality produces a low contrast image. Therefore, in the proposed methodology, first, we have enhanced the brain MR image by using histogram equalization techniques as illustrated in Fig. 5. Secondly, we have considered this enhanced MR brain image for further processing of removal of the skull from the brain part by using mathematical morphology techniques as illustrated in Fig. 6. It has been noted that the results of our presented methodology can be comparable to other [26 and 27] morphological based skull stripping methods stated in the literature.

A. Evaluation Metrics

The efficiency of image contrast enhancement has been measured using Peak signal to noise ratio, Signal to noise ratio, Mean absolute error, Root mean square error [48]. The performance of image contrast enhancement is depicted in Table I. While the equations have been illustrated.

$$PSNR = 10\log_{10}\left[\frac{\max(r(x,y))^2}{\frac{\sum_0^n x-1\sum_0^n y-1\left[r(x,y)-t(x,y)\right]^2}{nx.ny}}\right] \quad (6)$$

$$RMSE = \sqrt{\frac{\sum_0^n x-1\sum_0^n y-1\left[r(x,y)-t(x,y)\right]^2}{nx.ny}} \quad (7)$$

$$MAE = \frac{\sum_0^n x-1\sum_0^n y-1\,|\,r(x,y)-t(x,y)\,|}{nx.ny} \quad (8)$$

$$SNR = 10.\log_{10}\left[\frac{\sum_0^n x-1\sum_0^n y-1\left[r(x,y)\right]^2}{\sum_0^n x-1\sum_0^n y-1\left[r(x,y)-t(x,y)\right]^2}\right] \quad (9)$$

Fig. 5 to 9 illustrates different types of enhanced and respective skull stripping images using the proposed methodology.

In this study, the comparison between skull stripping and manually marked ground truth has been done using two standards such as Jaccard Coefficient and similarity coefficient Dice [49] as depicted in Table II. The proposed methodology has also been implemented on various MR brain image sequences as shown in Fig. 7 to 9.

$$JaccardCoefficient = \frac{A(S)\cap A(G)}{A(S)\cup A(G)} \quad (10)$$

$$DiceCoefficient = \frac{2\,|\,A(S)\cap A(G)\,|}{|\,A(S)\,|+|\,A(G)\,|} \quad (11)$$

It has been observed from the experimental results that the proposed methodology can be useful for different image analysis applications such as tumor classification, segmentation, and characterization.

TABLE. I. Performance of Contrast Enhancement Algorithm

	PSNR	RMSE	MAE	SNR
T2-W	25.1	12.6	5.9	22.9
T1-W	26.6	11.4	5.0	23.3
FLAIR	29.1	11.1	4.9	23.7
DW1	27.4	10.7	4.8	24.5
Avg. Perfor.	27.1	11.5	5.0	23.6

TABLE. II. Performance of Skull Stripping Algorithm

	Dice	Jaccard
T2-W	92.4	89.9
T1-W	93.0	90.1
FLAIR	95.6	91.3
DW1	96.1	92.1
Avg. Perfor.	94.3	90.8

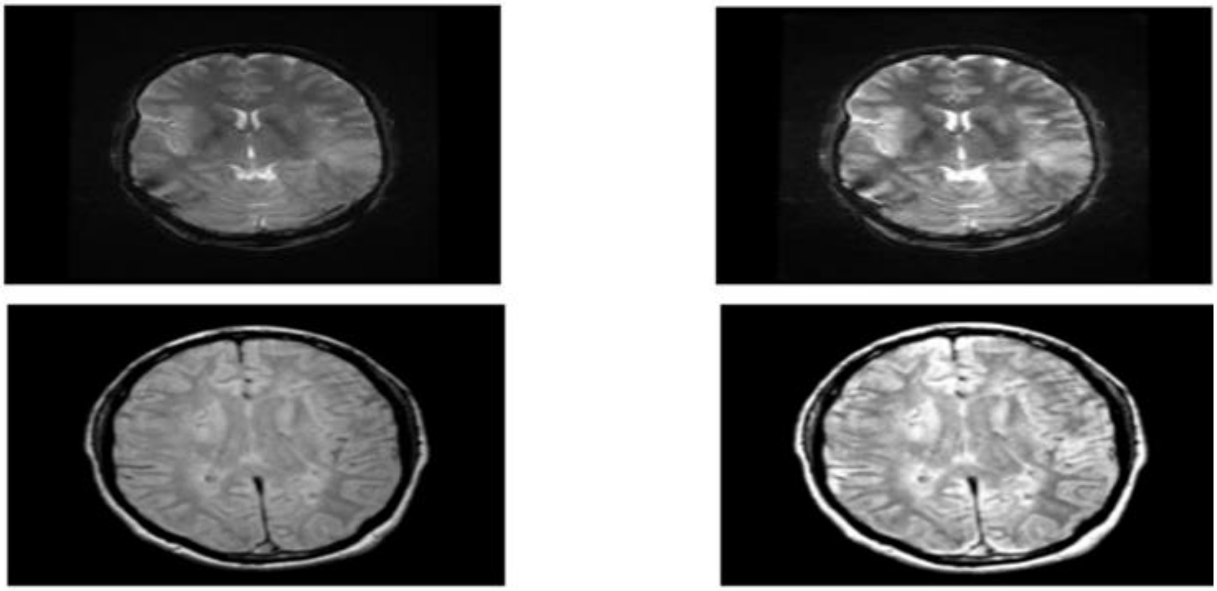


Fig. 5. Left Column Illustrates Original Input MR Brain Images while the Right Column Illustrates enhanced Output Images of MR Brain Images.

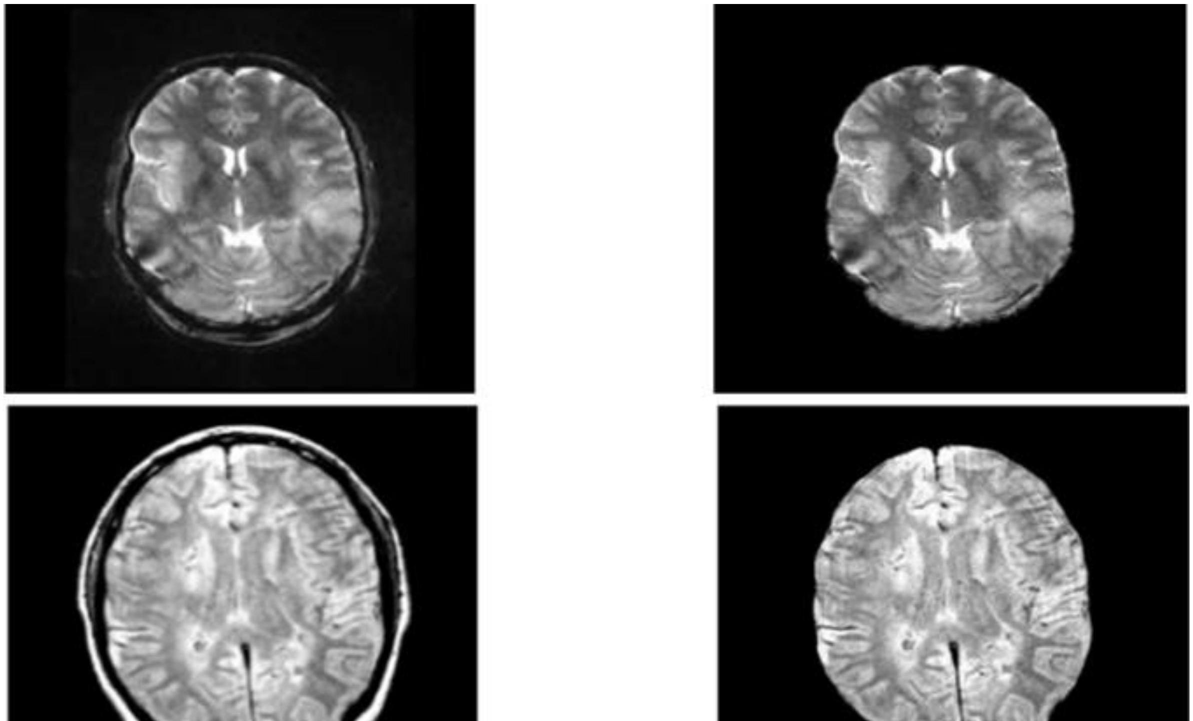


Fig. 6. Left Column Illustrates T2-W enhanced MR Images having Skull. While the Right Column Illustrates the Output of Skull Stripping Images.

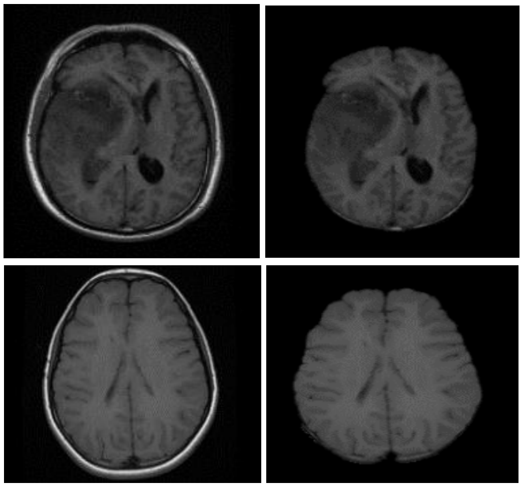


Fig. 7. Left Column Illustrates T1-W enhanced MR Images having Skull. While the Right Column Illustrates the Output of T1-W Skull Stripping Images.

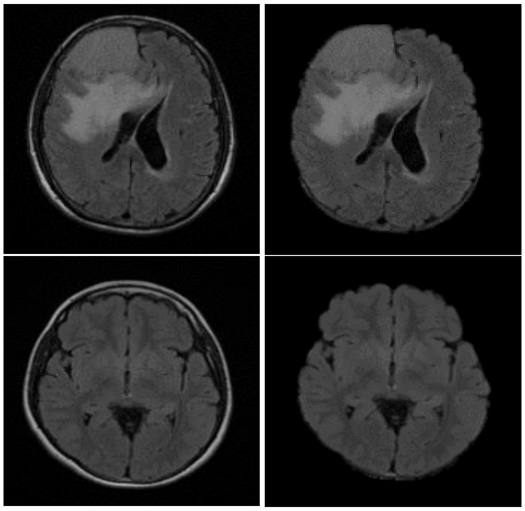


Fig. 8. Illustrates FLAIR enhanced Image and Respective Skull Stripped Image.

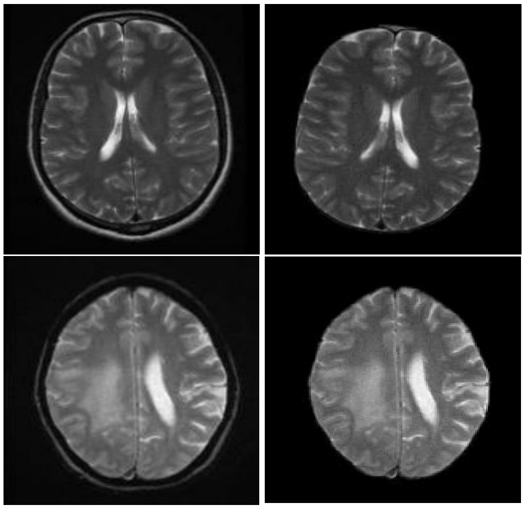


Fig. 9. Illustrates DW1 enhanced Image and Respective Skull Stripped Image.

VIII. Conclusion

In medical image segmentation applications, image preprocessing is an essential step to maximize classification accuracy. The images of MR imaging modality are low contrast and comprised of rician noise and salt and pepper noise. Therefore, these kinds of brain MR images are not helpful for physicians to diagnose a disease. To overcome this problem we have used histogram equalization techniques to enhance the brain MR images. However, the removal of the skull part from the brain part is also very helpful for the physicians to diagnose a disease accurately. So for skull stripping, we have used mathematical morphology techniques. This proposed algorithm works effectively on 2D MR brain images. It has been observed from the results that this methodology can be employed with many MR brain imaging applications and can be comparable to other morphological based skull stripping method.

IX. Future Work

In future work, we will focus on the solution of similar intensity segmentation of the intracranial and non-cerebral tissue of the brain. Also, this proposed study can be evolved for the preprocessing of 3D MR brain images.

X. Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

We are grateful to Department of Computer Engineering, Changwon National University for the immense support in piloting this work. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2018R1C1B6008187).

References

[1] H. K. Hahn and H. O. Peitgen, “The skull stripping problem in MRI solved by a single 3D watershed transform,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 1935, pp. 134–143, 2000.

[2] A. El-Baz, M. F. Casanova, G. Gimel’Farb, M. Mott, and A. E. Switwala, “A new image analysis approach for automatic classification of autistic brains,” 2007 4th IEEE Int. Symp. Biomed. Imaging From Nano to Macro - Proc., pp. 352–355, 2007.

[3] C.R. Guttmann, F.A. Jolesz, R. Kikinis, R.J. Killiany, M.B. Moss, T. Sandor, and M.S. Albert, “White matter changes with normal aging,” Neurology, vol. 50, issue-4, pp. 972–978, April, 1998.

[4] P. F. Buckley et al., “Three-dimensional magnetic resonance-based morphometrics and ventricular dysmorphology in schizophrenia,” Biol. Psychiatry, vol. 45, no. 1, pp. 62–67, 1999.

[5] E. Jackson, P. Narayana, J. Wolinsky and T. Doyle, “Accuracy and reproducibility in volumetric analysis of multiple,” Jou. Comput. Assisted Tomography, vol. 17, issue-2, pp. 200–205, March-April, 1993.

[6] K. Jafari-Khouzani, M.-R. Siadat, H. Soltanian-Zadeh, and K. Elisevich, “Texture analysis of hippocampus for epilepsy,” Med. Imaging 2003 Physiol. Funct. Methods, Syst. Appl., vol. 5031, p. 279, 2003.

[7] M. Stella Atkins and B. T. Mackiewich, “Fully automatic segmentation of the brain in MRI,” IEEE Trans. Med. Imaging, vol. 17, no. 1, pp. 98–107, 1998.

[8] R. W. . H. Cox J.S., “ AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages.,” Comput. Biomed. Res., vol. 29, no. 29, pp. 162–173, 1996.

[9] A. Kelemen, G. Székely, and G. Gerig, "Elastic model-based segmentation of 3-D neuroradiological data sets," IEEE Trans. Med. Imaging, vol. 18, no. 10, pp. 828–839, 1999.

[10] S. M. Smith, "Fast Robust Automated Brain Extraction," vol. 155, pp. 143–155, 2002.

[11] L. Lemieux, G. Hagemann, K. Krakow, and F. G. Woermann, "Fast, accurate, and reproducible automatic segmentation of the brain in T1-weighted volume MRI data," Magn. Reson. Med., vol. 42, no. 1, pp. 127–135, 1999.

[12] P. Maji and S. Roy, "Rough-fuzzy clustering and unsupervised feature selection for wavelet based MR image segmentation," PLoS One, vol. 10, no. 4, pp. 1–30, 2015.

[13] R. Adams and L. Bischof, "Seeded Region Growing," IEEE Trans. Pattern Anal. Mach. Intell., vol. 16, no. 6, pp. 641–647, 1994.

[14] H. K. Hahn and H. O. Peitgen, "The skull stripping problem in MRI solved by a single 3D watershed transform," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 1935, pp. 134–143, 2000.

[15] J. G. Park and C. Lee, "Skull stripping based on region growing for magnetic resonance brain images," Neuroimage, vol. 47, no. 4, pp. 1394–1407, 2009.

[16] S. F. Eskildsen et al., "BEaST: Brain extraction based on nonlocal segmentation technique," Neuroimage, vol. 59, no. 3, pp. 2362–2373, 2012.

[17] F. J. Galdames, F. Jaillet, and C. A. Perez, "An accurate skull stripping method based on simplex meshes and histogram analysis for magnetic resonance images," J. Neurosci. Methods, vol. 206, no. 2, pp. 103–119, 2012.

[18] R. A. Heckemann et al., "Brain extraction using label propagation and group agreement: Pincram," PLoS One, vol. 10, no. 7, pp. 1–18, 2015.

[19] J. E. Iglesias, C. Y. Liu, P. M. Thompson, and Z. Tu, "Robust brain extraction across datasets and comparison with publicly available methods," IEEE Trans. Med. Imaging, vol. 30, no. 9, pp. 1617–1634, 2011.

[20] J. E. Iglesias, C. Y. Liu, P. M. Thompson, and Z. Tu, "Robust brain extraction across datasets and comparison with publicly available methods," IEEE Trans. Med. Imaging, vol. 30, no. 9, pp. 1617–1634, 2011.

[21] S. Roy, J. A. Butman, and D. L. Pham, "Robust skull stripping using multiple MR image contrasts insensitive to pathology," Neuroimage, vol. 146, no. November, pp. 132–147, 2017.

[22] S. A. Sadananthan, W. Zheng, M. W. L. Chee, and V. Zagorodnov, "Skull stripping using graph cuts," Neuroimage, vol. 49, no. 1, pp. 225–239, 2010.

[23] F. Ségonne et al., "A hybrid approach to the skull stripping problem in MRI," Neuroimage, vol. 22, no. 3, pp. 1060–1075, 2004.

[24] D. W. Shattuck and R. M. Leahy, "Brainsuite: An automated cortical surface identification tool," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 1935, pp. 50–61, 2000.

[25] F. Ségonne et al., "A hybrid approach to the skull stripping problem in MRI," Neuroimage, vol. 22, no. 3, pp. 1060–1075, 2004.

[26] R. Roslan, N. Jamil, and R. Mahmud, "Skull stripping of MRI brain images using mathematical morphology," Proc. 2010 IEEE EMBS Conf. Biomed. Eng. Sci. IECBES 2010, no. December, pp. 26–31, 2010.

[27] S. Mohsin, S. Sajjad, Z. Malik, and A. H. Abdullah, "Efficient Way of Skull Stripping in MRI to Detect Brain Tumor by Applying Morphological Operations, after Detection of False Background," Int. J. Inf. Educ. Technol., no. July 2019, pp. 335–337, 2012.

[28] S. Roy, S. Nag, I. K. Maitra, P. Samir, and K. Bandyopadhyay, "Artefact Removal and Skull Elimination from MRI of Brain Image," no. June, 2014.

[29] J. E. Iglesias, C. Y. Liu, P. M. Thompson, and Z. Tu, "Robust brain extraction across datasets and comparison with publicly available methods," IEEE Trans. Med. Imaging, vol. 30, no. 9, pp. 1617–1634, 2011.

[30] S. F. Eskildsen et al., "BEaST: Brain extraction based on nonlocal segmentation technique," Neuroimage, vol. 59, no. 3, pp. 2362–2373, 2012.

[31] R. C. Gonzales and R. E. Woods, Digital Image Processing, Second Edition, Prentice Hall, 2002.

[32] "Harvard Medical School Data," www.med.harvard.edu/AANLIB

[33] R. M. Quencer and W. G. Bradley, "MR imaging of the brain: What constitutes the minimum acceptable capability?," Am. J. Neuroradiol., vol. 22, no. 8, pp. 1449–1450, 2001.

[34] M. Cheour, Advantages of brain MRI, 2010, Available at, RadiologyInfo.org.

[35] P. Schmid, "Segmentation of digitized dermatoscopic images by two-dimensional color clustering," IEEE Trans. Med. Imaging, vol. 18, no. 2, pp. 164–171, 1999.

[36] NLM-National Library of Medicine, (Rockville Pike, Bethesda U.S., 2011), Available online at: http://www.nlm.nih.gov.

[37] J. Kleesiek et al., "Deep MRI brain extraction: A 3D convolutional neural network for skull stripping," Neuroimage, vol. 129, pp. 460–469, 2016.

[38] G. Litjens et al., "A survey on deep learning in medical image analysis," Med. Image Anal., vol. 42, no. December 2012, pp. 60–88, 2017.

[39] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[40] P. Kalavathi and V. B. S. Prasath, "Methods on Skull Stripping of MRI Head Scan Images—a Review," J. Digit. Imaging, vol. 29, no. 3, pp. 365–379, 2016.

[41] S. S. Mohseni Salehi, D. Erdogmus, and A. Gholipour, "Auto-Context Convolutional Neural Network (Auto-Net) for Brain Extraction in Magnetic Resonance Imaging," IEEE Trans. Med. Imaging, vol. 36, no. 11, pp. 2319–2330, 2017.

[42] N. H. M. Duy, N. M. Duy, M. T. N. Truong, P. T. Bao, and N. T. Binh, "Accurate brain extraction using Active Shape Model and Convolutional Neural Networks," no. February, 2018.

[43] R. Dey, Y. Hong, "CompNet: Complementary Segmentation Network for Brain MRI Extraction," arXiv 2018, preprint arXiv:1804.00521.

[44] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: Learning dense volumetric segmentation from sparse annotation," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9901 LNCS, no. June 2016, pp. 424–432, 2016.

[45] B. Puccio, J. P. Pooley, J. S. Pellman, E. C. Taverna, and R. C. Craddock, "The preprocessed connectomes project repository of manually corrected skull-stripped t1-weighted anatomical mri data," Gigascience, vol. 5, no. 1, pp. 1–7, 2016.

[46] Fischl, B. FreeSurfer. Neuroimage 2012, 62, 774–781, doi:10.1016/j.neuroimage.2012.01.021.

[47] A. Fedorov, J. Johnson, E. Damaraju, A. Ozerin, V. Calhoun, and S. Plis, "End-to-end learning of brain tissue segmentation from imperfect labeling," Proc. Int. Jt. Conf. Neural Networks, vol. 2017-May, pp. 3785–3792, 2017.

[48] K. J. Jang et al. "Measuremnet of Image Quality in CT Images Reconstructed with Different Kernels," Jour. Of the Korean Phy. Soc, 58(2), pp. 334-342, 2011.

[49] Jaccard, P.: The Distribution of Flora in Alpine Zone, New Phytol, 11(2), 1912, 37-50.

AUTHOR'S PROFILE

https://orcid.org/0000-0002-0184-7620

Zahid Ullah, is currently doing Ph.D in Computer Engineering from Changwon National University South Korea. He received MS in Computer Science from SZABIST, Islamabad, Pakistan in 2015, and BS in Information Technology from University of Malakand Pakistan in 2011. His area of interest are Image Processing, Medical Imaging, Computer Vision, and Machine Learning Techniques.

https//orcid.org/0000-0001-6966-1569

Dr. Su-Hyun, Lee, is working as a Professor and Head of department of Computer Engineering, at Changwon National University South Korea. He is working in multiple desciplines, especially Bio-Informatics, Algorithm, and Programming.

Dr. Donghyeok An received the B.S. degree in computer science from Handong University, Pohang, South Korea, in 2006, and the Ph.D. degree in Computer Science from Korea Advanced Institute of Science and Technology (KAIST) in Feb. 2013. He was a post doctor and a visiting professor at Sungkyunkwan University from Mar. 2013 to Feb. 2014. He was a senior engineer at Samsung Electronics from 2014 to 2015. He was an assistant professor in the Department of Computer Engineering, Keimyung University from 2015 to 2017. He is currently an assistant professor in the Department of Computer Engineering, Changwon National University, Changwon, South Korea. His research interests include 5G networks, ultra reliable low latency communication, and Internet of Things.

Analysis of Web Content Quality Factors for Massive Open Online Course using the Rasch Model

Wan Nurhayati Wan Ab Rahman[1], Hazura Zulzalil[2], Iskandar Ishak[3], Ahmad Wiraputra Selamat[4]
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia, Serdang, Malaysia

***Abstract*—The lack of understanding among content providers towards the quality of MOOC motivates the development of several MOOC quality models. However, none was focused on the web content from the perspective of content providers or experts despite the facts that their views are important particularly in the development phase. MOOCs learners and instructors definitely understand the functional external quality, but content providers have better understanding to the internal qualities, which is required during the development phase. The initial quality model for MOOC web content based on 7C's of Learning Design and PDCA model for continuity have been proposed, consisted of nine categories and 54 factors. This research focuses on the validation towards the proposed model by content providers and experts to provide systematic evidence of construct validity. This involved two main processes; content validity test and survey on acceptability. The content validity test was conducted to confirm the agreeability of proposed categories and factors among respondents. The Dichotomous Rasch model was utilized to explain the conditional probability of a binary outcome, given the person's agreeability level and the item's endorsability level. Subsequently, the survey on acceptability was conducted to obtain confirmation and verification from the experts group pertaining on MOOC web content quality factors. Rasch Rating Scale model was used since it specifies the set of items, which share the same rating scale structure. The usage of the Rasch Model in instrument development generally ease variable measurement by converting the nonlinear raw data to linear scale, while assists researchers in tackling fitness validation and other instrumentation issues like person reliability and unidimensionality. This paper demonstrates the strengths of applying Rasch Model in construct validation and instrument building, which provides a strong foundation for the model adaptation as a methodological tool.**

***Keywords*—Web content; quality model; hierarchical model; Rasch Model; rating scale; survey reliability**

I. Introduction

Widespread acceptance among instructors and learners since its introduction in 2008 does not prevent the Massive Open Online Course (MOOC) from receiving a number of criticism in its implementation. Some of the major issues is pertaining on its web content weaknesses, despite its importance in maintaining learner's engagement as supported by [1]. To overcome this, previous research by [2] have proposed a web content quality model for MOOC from the perspective of content providers, which takes into consideration the aspects of external and internal quality. The model intents to facilitate the understanding of content providers into the right facet of producing a quality web content for MOOC.

The quality factor is an instrument that needs to be empirically verified to ensure its reliability and usefulness in the real-world environment. It leads to the main objective of this research which to validate the proposed web content quality model for MOOC by [2] and its definitions from the perspective of content providers and MOOC experts. In order to achieve this objective, two tests were conducted: content validity test and survey on acceptability. This test meant to fulfill six criteria of construct validity proposed by [3] which is content, substantive, structure, generalizability, external and consequential.

The Rasch Model was implemented due to its capability to assess the construct validity by transforming the ordinal data into a linear score, before it's been evaluated through the use of parametric statistical tests as proven by [4]. This analysis method also enables researchers to make critical corrections to the raw test score by implementing fitness validation. It is being utilized by number of instrument validation such as blog quality model by [5] and customer satisfaction for service quality by [6]. Moreover, the web content quality model is developed within the hierarchical factor-criteria-metrics (FCM) framework, similar with several hierarchical models like McCall, Boehm and ISO/IEC 9126. Therefore, it is important to ensure that all quality items have single dimension towards the model objective, or called unidimensionality. The Rasch Model was used to ensure the unidimensionality compliance through the function of Category Probability Modes and Principal Component Analysis [7]. Its adaptation along with data fitness validation and the probability of an item to be accepted is explained in this research.

The rest of this paper is organized as follows: Section II describes briefly about the development of Web Content Quality Model for MOOC and the Rasch Model method; Section III explains how the content validity and survey on acceptability were conducted; Section IV discusses the results and discussions; and finally, Section V touches on the conclusions.

II. Literature Review

A. The Development of Web Content Quality Model for MOOC

The initial web content quality model for MOOC as reference to content providers has been proposed by [2] as depicted in Fig. 1. Its development began with the determination of quality factors through the process of content analysis, which involved three activities: (i) Review the existing and possible quality factors from online library in duration of 2010 to 2018 (ii) Combining the set of factors to cross check any redundancies as applied by [8] and (iii) Assigning the factors into respective categories. The content analysis yields 54 quality factors, which assigned into nine categories that modified and customized from the 7C's model.

Author in [9] point out that quality evaluation by untrained or end users is questionable and not comprehensive. Therefore, instrument validation from the perspective of content providers and MOOC experts was applied in this research to secure the validity of the proposed quality factors, as acknowledged by number of researchers like [10] and [11]. Content validity is an important procedure in scale development, which the degree of an instrument has appropriate sample of items for the constructs that being measured [12]. This test also is a non-statistical type of validity that involves systematic examination of the survey content to determine whether it covers a representative sample of the behavior domain to be measured. Its main objective is to ensure that the instruments represent all facets of a given constructs, as well as providing a solid basis for rigorous validation evaluation [13].

The survey on acceptability measure the level of acceptability among respondents to the proposed categories and factors of a model based on the steps proposed by [13] such as survey planning, availability of the resource, survey design, data collection planning and selection of participants. The survey can be executed through a structured standardized interview that follow determined and specific questionnaire. This data collection methodology has been applied by number of research such as [5] and [14] to validate the newly developed model.

B. The Rasch Model

Rasch measurement model which introduced by a Danish mathematician named Georg Rasch in 1960 is a psychometric technique to improve researchers construct instruments precision, monitor instrument quality, and compute respondents' performances [15]. It creates measurements from categorical data such as questionnaire responses, as a function of the trade-off between the respondent's abilities and item difficulties [16]. Rasch model also enables researcher to make critical corrections when using raw test score or survey data. The mathematical theory underlying the Rasch models is a special case of item response theory and generalized linear model.

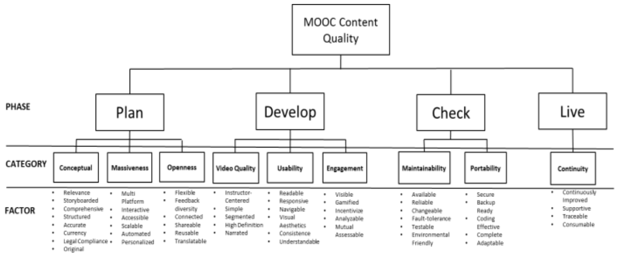


Fig. 1. Initial Web Content Quality Model for MOOC Proposed by [2].

Rasch Dichotomous Measurement Model is a probabilistic model which considers two aspects: (i) Difficulty of the item (ii) Ability of respondent to verify the item. The model explains the conditional probability of a binary outcome (in this research, agree or disagree), given the person's agreeability and the item's endorsability level. It is based on the logic that all respondents have a higher probability of answering easier items and a lower probability of answering difficult items. This is expressed mathematically as:

$$\Pr\{X_{ni} = 1\} = \frac{e^{\beta n-\ \delta_i}}{1+e^{\beta n-\ \delta_i}} \tag{1}$$

$\Pr\{X_{ni} = 1\}$ refers to the probability of agreement of person n towards the item i, while β_n is the ability of person n and δ_i is the difficulty of item i. Thus, in the case of a dichotomous attainment item, it is shown that the log odds or logits of correct response by a person to an item is equal to $\beta_n - \delta_i$. The reason is based on the need to transform it to logits in order to obtain a linear interval scale [4]. Given two examinees with different ability parameters β_1 and β_2 and an arbitrary item with difficulty δ_i, compute the difference in logits for these two examinees by $(\beta_1 - \delta_i)$ - $(\beta_2 - \delta_i)$. This difference becomes $\beta_1 - \beta_2$. The logistic function in the equation as it allows for making estimates of β_n and δ_i independently each other. Hence, the estimates β_n are independent of the effect of δ_i and the estimates of δ_i are independent of the effect of β_n. The separation between these two parameters provides a simple yet powerful model to assess survey response, making it possible to obtain a linear scale and generalized measurement [7]. Constant e is referring to natural log function (2.7183) of the difference between person's ability and item's difficulty. This can be expressed mathematically as follows:

$$\ln\left(\frac{\frac{e^{(\beta n-\delta_i)}}{1-e^{(\beta n-\delta_i)}}}{1-\left(\frac{e^{(\beta n-\delta_i)}}{1-e^{(\beta n-\delta_i)}}\right)}\right) = \beta_n - \delta_i \tag{2}$$

A direct comparison between person's ability and an item's difficulty can be obtained as follows: the probability of success on any item, given's a person's ability and item's difficulty. It is divided by the probability of failure on any item and the natural log of resulting expression that provides the comparison [7]. This implies that persons and items can be compared directly as the characteristics of both have been separated. This unique property is called parameter separation [16].

Rasch Rating Scale Model is the extension of the Rasch Dichotomous Model. It derived from concept of threshold as the item is modelled of having three threshold if it contains four response choices. Every item threshold labelled with k has its own difficulty estimation F, and this is modelled as the threshold at which a person has an equal probability of choosing one category over another. For example, the first threshold is modelled as the probability of choosing a response of "2" (disagree) over the response of "1" (strongly disagree), which then estimated using the formula as follows:

$$P_{ni1}\{X_{ni} = 1/B_n, D_i, F_1\} = \frac{e^{(Bn-[D_i+F_1])}}{1+e^{(Bn-[D_i+F_1])}} \tag{3}$$

P_{ni1} is the probability of person n in choosing "Disagree" (Category 2) over "Strongly Disagree" (Category 1) on any item i. In this equation, F_1 is the difficulty of the first threshold, and this difficulty calibration is estimated only once for this threshold across the entire set of items in the rating scale.

III. Methodology

A. Content Validity Test

The content validity was conducted to confirm whether the content is agreeable to the respondent, hence provides the empirical evidence of content aspect in construct validity. Reference [3] explained that besides content, the aspect of consequential, substantive, structural, external and generalizability are also contribute to the construct validity. In this research, the content validity test was conducted through web-based online survey in order to ease data gathering, increase response rate, minimize cost and automate data input as supported by [17]. Google Forms was utilized as survey instrument based on its advantages like high reachability, freely available, easy to use and automatic data response input [18].

Participants were invited via communication tools like e-mail, Facebook, Twitter and MOOC platforms to complete the online survey. They were selected openly through profiling processes with the assistance from MOOC community like the Malaysia E-Learning Higher Learning Institution Coordinator (MEIPTA) and The Australasian Council on Open, Distance and e-Learning (ACODE). The fit respondents also selected from professional sites like LinkedIn, authors of paper that used in literature review and experts from any related conference or workshops. The respondent resume and experiences were examined through their profiles available in their websites to gauge their knowledge and expertise on MOOC.

Rasch Dichotomous Measurement Model has been adapted as the analytical method to explain the conditional probability of the binary outcome, which is agree or disagree. The questionnaire data was setup in a free Rasch analysis application called Bond&FoxStep, which is the customized version of proprietary Winstep®. Through this application, analysis of reliability, person separation and principal component were carried out. Measurement of acceptance level for items and persons was made through the Wright Map, while measurement of scale was executed through Rating (Partial Credit) Scale.

B. Survey on Acceptability

The survey on acceptability was conducted after the content validity test to obtain confirmation and verification from the content providers and experts concerning the web content quality factors for MOOC. The survey was executed through structured standardised interview in order to get the optimum results. The content providers and experts were selected mostly from the higher learning institution and MOOC platform developer**.** The survey consists close-ended and open-ended questions to gain variety of recommendations and comments.

Before the implementation of survey on acceptability, the pretesting was conducted on the redesigned questionnaire to assess its clarity, readability and understandability to the participants. This process involved four field experts comprising statistician, MOOC expert, language expert and web designer expert. Once all of them were satisfied, the reviewed questionnaire was distributed to 49 MOOC experts and content providers. The questionnaire comprised of two parts: (i) Part I: The respondent states their gender, age and occupation. (ii) Part II: The respondent indicates the extent on which they agreed or disagreed with the proposed MOOC quality content on the scale of 1 to 5 (1 – strongly disagree and 5 – strongly agree). An open question was also included to draw further recommendations and comments.

Similar to the content validity test, the survey analysis was executed through Bond&Fox application. Data was tabulated and analysed using Rasch Rating Scale Model, given that the survey deal with multiple response category item. Rasch Rating Scale can deal with a small sample size of 50 to provide useful and reliable estimates for item calibrations, at a 99% Confidence Interval or within ± 1 logits [16].

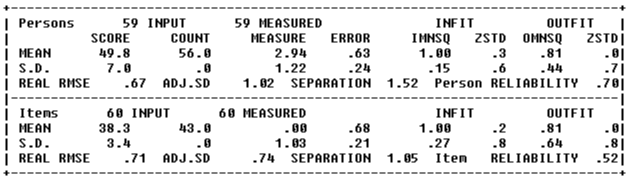
IV. Result and Discussion

A. Content Validity Test

Fig. 2 depicts the summary statistics for the sample of 59 person on the 60 dichotomous scale items, comprising of 9 categories and 51 quality factors. The mean of the person measures is 2.94 (SE .63) that is higher than the 0 calibration of the item scale, indicates that majority of respondents found this questionnaire relatively understandable. The summary statistics for item and person imply satisfactory fit to the model. The value of person reliability which is higher than .67 (at 95% confidence level) means the test discriminate the sample into enough level, indicates the instruments for measuring content validity is reliable for measurement purpose. The item reliability which is .52 (at 95% confidence level) has no traditional equivalent and can be ignored for this purpose.

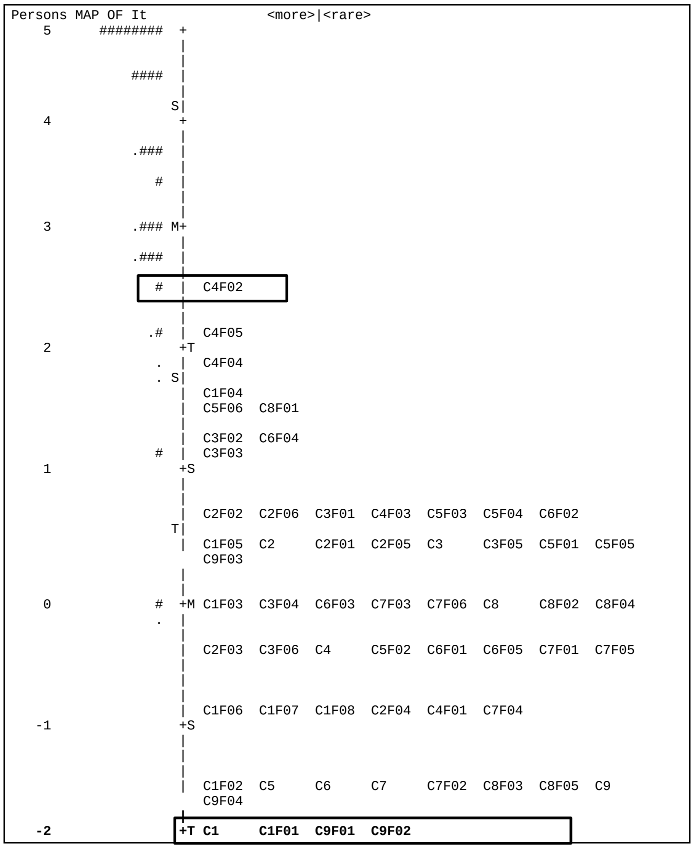
The Wright Map in Fig. 3 shows the distribution of person on the left and the item agreement on the right, represented by category ID and factor ID. The agreeability level of person are clearly shown on the map, as the most agreeable items like C1 (Conceptual), C1F01 (Relevance), C9F01 (Consumable) and C9F02 (Continuous Improvement) are that located at -2.90 logits (SE 1.84). On the contrary, the least agreeable items which is C4F02 (Instructor-Centred) located on top of the item distribution at +2.46 logits (SE 0.35). The mean of person distribution μ_{person}=+2.94 logit is higher than the mean of the item distribution μ_{person}=0.00 logits, indicates that most of the respondents involved in the content validity test have tendency of agreeing the proposed categories and assigned factors definition. The probability of person's agreement with the identified categories and factors were calculated using (1). With the mean of 2.94, respondents generally indicate their level of agreement at 94.97%, which is above the 70% threshold limit of Cronbach's Alpha as shown in the following calculation:

$$\Pr\{X_{ni} = 1\} = \frac{e^{2.94-0}}{1+e^{2.94-0}} = 94.97\%$$



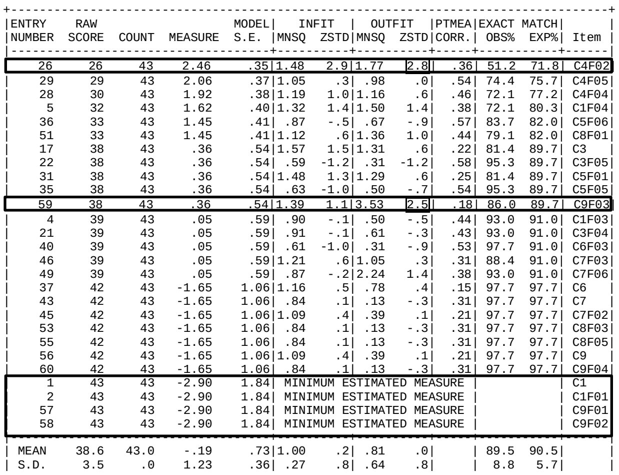
```
+-----------------------------------------------------------------------------+
| Persons      59 INPUT      59 MEASURED              INFIT         OUTFIT    |
|            SCORE      COUNT     MEASURE   ERROR     IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN        49.8       56.0        2.94     .63      1.00     .3    .81     .0|
| S.D.         7.0         .0        1.22     .24       .15     .6    .44     .7|
| REAL RMSE    .67  ADJ.SD     1.02  SEPARATION  1.52  Person RELIABILITY  .70|
|-----------------------------------------------------------------------------|
| Items       60 INPUT      60 MEASURED              INFIT         OUTFIT    |
| MEAN        38.3       43.0         .00     .68      1.00     .2    .81     .0|
| S.D.         3.4         .0        1.03     .21       .27     .8    .64     .8|
| REAL RMSE    .71  ADJ.SD      .74  SEPARATION  1.05  Item   RELIABILITY  .52|
+-----------------------------------------------------------------------------+
```

Fig. 2. Summary Statistics of the Content Validity Test.



```
Persons MAP OF It              <more>|<rare>
    5      ########  +
                     |
                     |
               ####  |
                     |
                    S|
    4                +
                     |
               .###  |
                     |
                  #  |
                     |
                     |
    3          .### M+
                     |
               .###  |
                     |
                  #  |  C4F02
                     |
                     |
                 .#  |  C4F05
    2                +T
                  .  |  C4F04
                  . S|
                     |  C1F04
                     |  C5F06  C8F01
                     |
                     |  C3F02  C6F04
                  #  |  C3F03
    1                +S
                     |
                     |
                     |  C2F02  C2F06  C3F01  C4F03  C5F03  C5F04  C6F02
                    T|
                     |  C1F05  C2     C2F01  C2F05  C3     C3F05  C5F01  C5F05
                        C9F03
                     |
                     |
    0             #  +M C1F03  C3F04  C6F03  C7F03  C7F06  C8     C8F02  C8F04
                  .  |
                     |
                     |  C2F03  C3F06  C4     C5F02  C6F01  C6F05  C7F01  C7F05
                     |
                     |
                     |
                     |  C1F06  C1F07  C1F08  C2F04  C4F01  C7F04
   -1                +S
                     |
                     |
                     |
                     |  C1F02  C5     C6     C7     C7F02  C8F03  C8F05  C9
                        C9F04
                     |
   -2                +T C1     C1F01  C9F01  C9F02
```

Fig. 3. The Wright Map of the Content Validity Test.



```
+------------------------------------------------------------------------------------+
|ENTRY    RAW                   MODEL|   INFIT  |  OUTFIT  |PTMEA|EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.| OBS%  EXP%| Item |
|------------------------------------+----------+----------+-----+-----------+------|
|    26     26     43    2.46     .35|1.48   2.9|1.77   2.8|  .36| 51.2  71.8| C4F02|
|    29     29     43    2.06     .37|1.05    .3| .98    .0|  .54| 74.4  75.7| C4F05|
|    28     30     43    1.92     .38|1.19   1.0|1.16    .6|  .46| 72.1  77.2| C4F04|
|     5     32     43    1.62     .40|1.32   1.4|1.50   1.4|  .38| 72.1  80.3| C1F04|
|    36     33     43    1.45     .41| .87   -.5| .67   -.9|  .57| 83.7  82.0| C5F06|
|    51     33     43    1.45     .41|1.12    .6|1.36   1.0|  .44| 79.1  82.0| C8F01|
|    17     38     43     .36     .54|1.57   1.5|1.31    .6|  .22| 81.4  89.7| C3   |
|    22     38     43     .36     .54| .59  -1.2| .31  -1.2|  .58| 95.3  89.7| C3F05|
|    31     38     43     .36     .54|1.48   1.3|1.29    .6|  .25| 81.4  89.7| C5F01|
|    35     38     43     .36     .54| .63  -1.0| .50   -.7|  .54| 95.3  89.7| C5F05|
|    59     38     43     .36     .54|1.39   1.1|3.53   2.5|  .18| 86.0  89.7| C9F03|
|     4     39     43     .05     .59| .90   -.1| .50   -.5|  .44| 93.0  91.0| C1F03|
|    21     39     43     .05     .59| .91   -.1| .61   -.3|  .43| 93.0  91.0| C3F04|
|    40     39     43     .05     .59| .61  -1.0| .31   -.9|  .53| 97.7  91.0| C6F03|
|    46     39     43     .05     .59|1.21    .6|1.05    .3|  .31| 88.4  91.0| C7F03|
|    49     39     43     .05     .59| .87   -.2|2.24   1.4|  .38| 93.0  91.0| C7F06|
|    37     42     43   -1.65    1.06|1.16    .5| .78    .4|  .15| 97.7  97.7| C6   |
|    43     42     43   -1.65    1.06| .84    .1| .13   -.3|  .31| 97.7  97.7| C7   |
|    45     42     43   -1.65    1.06|1.09    .4| .39    .1|  .21| 97.7  97.7| C7F02|
|    53     42     43   -1.65    1.06| .84    .1| .13   -.3|  .31| 97.7  97.7| C8F03|
|    55     42     43   -1.65    1.06| .84    .1| .13   -.3|  .31| 97.7  97.7| C8F05|
|    56     42     43   -1.65    1.06|1.09    .4| .39    .1|  .21| 97.7  97.7| C9   |
|    60     42     43   -1.65    1.06| .84    .1| .13   -.3|  .31| 97.7  97.7| C9F04|
|     1     43     43   -2.90    1.84| MINIMUM ESTIMATED MEASURE |           | C1   |
|     2     43     43   -2.90    1.84| MINIMUM ESTIMATED MEASURE |           | C1F01|
|    57     43     43   -2.90    1.84| MINIMUM ESTIMATED MEASURE |           | C9F01|
|    58     43     43   -2.90    1.84| MINIMUM ESTIMATED MEASURE |           | C9F02|
|------------------------------------+----------+----------+-----+-----------+------|
| MEAN    38.6   43.0     -.19     .73|1.00    .2| .81    .0|     | 89.5  90.5|      |
| S.D.     3.5     .0     1.23     .36| .27    .8| .64    .8|     |  8.8   5.7|      |
```

Fig. 4. Item Measure for Content Validity Test.

Fig. 4 shows the item statistics that details the location of all items in Wright Map, as the top-most and bottom-most items on are equivalence. The fit statistics indicates that person fully agree with four estimated items which are C1, C1F01, C9F01 and C9F02. These items were retained in this analysis as it did not influence the measurement. In the context of Rasch analysis, infit and outfit determine the fitness of model accurately and indicate whether the item need to be deleted, rescored, or reworded. The item's infit / outfit mean square (MNSQ) value that falls outside the range of 0.6 to 1.4 and infit / outfit ZSTD value that fall outside -2.0 and +2.0 behaved more erratic than expected. The analysis performed on Outfit MNSQ and Outfit ZSTD columns reveals that all item adequately fit the model except C4F02 (*Instructor-Centred*) and C9F03 (*Traceable*). The ZSTD of C4F02 is 2.8 and C9F03 is 2.5, which considered misfits.

Crosschecking on the Guttman Scalogram as shown in Fig. 5 indicates that both misfit items, which are *Instructor-Centred* and *Traceable* have been underrated by a several person. For example, the person with ID F02 disagrees with *Instructor-Centred,* while most of the top is agree. That case is similar with the persons with ID F05 and F06 that disagree with *Traceable,* while the patterns of other person agree with it. This may due to the carelessness by the persons in attempting their work. However after verifying their MNSQ infit value which is within productive range (1.48 and 1.39, the range is within 0.4 to 1.6), the two misfits were validated. This criterion-reference interpretation of measure supports the technical quality of the content aspect in construct validity.

As stated in the content validity test objectives, two different aspects were analyzed: (i) the definition of categories and factors, and (ii) the assigning of factors into its respective categories. The probability of both aforementioned aspects was calculated based on logits measure. This also determines the revision's requirement for respondent's views from open-ended question. The formula of (1) was used to measure the probability for each categories. A threshold of 70% was set in line with the standard threshold limit of Cronbach Alpha [4]. It was then interpreted as follows:

a) Definition of categories and factors with probability to be agreed more than or equal to 70% will be accepted without any revision.

b) Definition of categories and factors with probability to be agreed less than 70% will be reviewed if related comments are provided by the respondents. The categories will be subsequently redefined whereas the factors will be discarded or amended if applicable.

For example, for the category C01 *Conceptual* that the value of person measure is 2.94 and item measure is -2.9, the calculation of probabilities is as follows:

$$P(\theta)\ \% = \beta n - \delta i$$

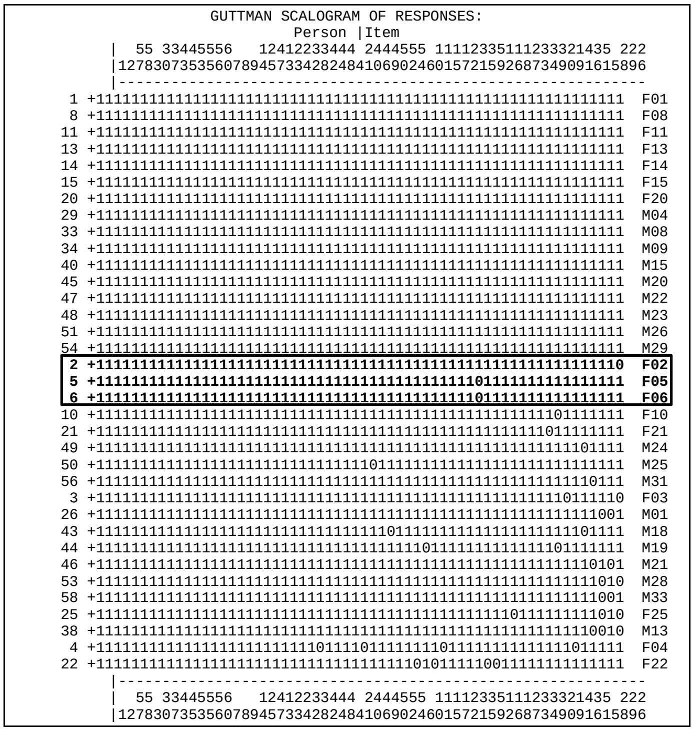
$$= 2.94 - (-2.55)$$

$$= 5.49$$

$$= \frac{e^{\beta n - \delta i}}{1 + e^{\beta n - \delta i}}$$

$$= \frac{e^{5.49}}{1 + e^{5.49}}$$

$$= 99.6\%$$



```
                      GUTTMAN SCALOGRAM OF RESPONSES:
                              Person |Item
    |  55 33445556   12412233444 2444555 11112335111233321435 222
    |127830735356078945733428248410690246015721592687349091615896
    |------------------------------------------------------------
  1 +111111111111111111111111111111111111111111111111111111111111  F01
  8 +111111111111111111111111111111111111111111111111111111111111  F08
 11 +111111111111111111111111111111111111111111111111111111111111  F11
 13 +111111111111111111111111111111111111111111111111111111111111  F13
 14 +111111111111111111111111111111111111111111111111111111111111  F14
 15 +111111111111111111111111111111111111111111111111111111111111  F15
 20 +111111111111111111111111111111111111111111111111111111111111  F20
 29 +111111111111111111111111111111111111111111111111111111111111  M04
 33 +111111111111111111111111111111111111111111111111111111111111  M08
 34 +111111111111111111111111111111111111111111111111111111111111  M09
 40 +111111111111111111111111111111111111111111111111111111111111  M15
 45 +111111111111111111111111111111111111111111111111111111111111  M20
 47 +111111111111111111111111111111111111111111111111111111111111  M22
 48 +111111111111111111111111111111111111111111111111111111111111  M23
 51 +111111111111111111111111111111111111111111111111111111111111  M26
 54 +111111111111111111111111111111111111111111111111111111111111  M29
  2 +111111111111111111111111111111111111111111111111111111111110  F02
  5 +111111111111111111111111111111111111111111101111111111111111  F05
  6 +111111111111111111111111111111111111111111101111111111111111  F06
 10 +111111111111111111111111111111111111111111111111111101111111  F10
 21 +111111111111111111111111111111111111111111111111111011111111  F21
 49 +111111111111111111111111111111111111111111111111111111101111  M24
 50 +111111111111111111111111111111101111111111111111111111111111  M25
 56 +111111111111111111111111111111111111111111111111111111110111  M31
  3 +111111111111111111111111111111111111111111111111111110111110  F03
 26 +111111111111111111111111111111111111111111111111111111111001  M01
 43 +111111111111111111111111111111111011111111111111111111101111  M18
 44 +111111111111111111111111111111111111101111111111111101111111  M19
 46 +111111111111111111111111111111111111111111111111111111110101  M21
 53 +111111111111111111111111111111111111111111111111111111111010  M28
 58 +111111111111111111111111111111111111111111111111111111111001  M33
 25 +111111111111111111111111111111111111111111111101111111111010  F25
 38 +111111111111111111111111111111111111111111111111111111110010  M13
  4 +111111111111111111111111101111011111111011111111111111011111  F04
 22 +111111111111111111111111111111111111010111110011111111111111  F22
    |------------------------------------------------------------
    |  55 33445556   12412233444 2444555 11112335111233321435 222
    |127830735356078945733428248410690246015721592687349091615896
```

Fig. 5. Guttman Scalogram of Content Validity Test.

The probability of agreement for item C01 *Conceptual* is 99.6%, which is higher than the set threshold of 70%. Therefore, *Conceptual* is accepted as one of the categories that form the web content quality model for MOOC. The results for the rest of the categories along with *Conceptual* are presented in Table I. It concludes that all nine proposed categories were agreed by the respondents along with its definitions, with the probability is between 91.8% and 99.6%.

The finding of assigning factors into respective categories is shown on Table II. It can be seen that 50 factors (with probability to be agreed more than 70%) remain in their respective categories. Based on these findings, only one factor which is *Instructor Centred* from *Video Quality* category has possibility of acceptance lower than 70%, to be exact 61.77%. Therefore, the definition needs to be reviewed. Table III shows the comments from respondents related to this factor.

TABLE. I. PROBABILITY TO BE AGREED FOR THE DEFINITIONS OF CATEGORY BY RESPONDENT

Code	Category Name	Person Measure	Item Measure	P(Θ) %
C01	Conceptual	2.97	-2.55	99.6
C02	Massiveness	2.97	0.15	94.4
C03	Openness	2.97	0.56	91.8
C04	Video Quality	2.97	-0.4	96.7
C05	Usability	2.97	-1.24	98.5
C06	Engagement	2.97	-1.24	98.5
C07	Maintainability	2.97	-1.24	98.5
C08	Portability	2.97	-1.24	98.5
C09	Continuity	2.97	-2.55	99.6

TABLE. II. PROBABILITY TO BE AGREED FOR THE DEFINITIONS OF QUALITY FACTORS BY RESPONDENT

Code	Factor Name	Person Measure	Item Measure	P(Θ) %
C1F01	Relevance	2.94	-2.9	99.71
C1F02	Currency	2.94	-1.65	98.99
C1F03	Legal Compliance	2.94	0.05	94.74
C1F04	Original	2.94	1.62	78.92
C1F05	Storyboarded	2.94	0.36	92.96
C1F06	Comprehensive	2.94	-0.85	97.79
C1F07	Structured	2.94	-0.85	97.79
C1F08	Accurate	2.94	-0.85	97.79
C2F01	Multi-Platform	2.94	0.36	92.96
C2F02	Scalable	2.94	0.63	90.97
C2F03	Personalized	2.94	-0.34	96.37
C2F04	Interactive	2.94	-0.85	97.79
C2F05	Automated	2.94	0.36	92.96
C2F06	Accessible	2.94	0.63	90.97
C3F01	Shareable	2.94	0.63	90.97
C3F02	Reusable	2.94	1.27	84.16
C3F03	Translatable	2.94	1.08	86.53
C3F04	Connected	2.94	0.05	94.74
C3F05	Feedback diversity	2.94	0.36	92.96
C3F06	Flexible	2.94	-0.34	96.37
C4F01	Segmented	2.94	-0.85	97.79
C4F02	Instructor-Centered	2.94	2.46	61.77
C4F03	Simple	2.94	0.63	90.97
C4F04	High Definition	2.94	1.92	73.50
C4F05	Narrated	2.94	2.06	70.68
C5F01	Navigable	2.94	0.36	92.96
C5F02	Readable	2.94	-0.34	96.37
C5F03	Understandable	2.94	0.63	90.97
C5F04	Visual Aesthetics	2.94	0.63	90.97
C5F05	Consistence	2.94	0.36	92.96
C5F06	Responsive	2.94	1.45	81.61
C6F01	Analyzable	2.94	-0.34	96.37
C6F02	Mutual Assessable	2.94	0.63	90.97
C6F03	Incentivize	2.94	0.05	94.74
C6F04	Gamified	2.94	1.27	84.16
C6F05	Visible	2.94	-0.34	96.37
C7F01	Changeable	2.94	-0.34	96.37
C7F02	Available	2.94	-1.65	98.99
C7F03	Fault tolerance	2.94	0.05	94.74
C7F04	Reliable	2.94	-0.85	97.79
C7F05	Testable	2.94	-0.34	96.37
C7F06	Environmental Friendly	2.94	0.05	94.74
C8F01	Coding Effective	2.94	1.45	81.61
C8F02	Complete	2.94	0.05	94.74
C8F03	Secure	2.94	-1.65	98.99
C8F04	Backup ready	2.94	0.05	94.74
C8F05	Adaptive	2.94	-1.65	98.99
C9F01	Consumable	2.94	-2.9	99.71
C9F02	Continuous Improvement	2.94	-2.9	99.71
C9F03	Traceable	2.94	0.36	92.96
C9F04	Supportive	2.94	-1.65	98.99

TABLE. III. RESPONDENTS' COMMENTS ON QUALITY FACTORS OF INSTRUCTOR-CENTRED

Quality Factors	**Comments from respondents**
Instructor-Centred	- Should be learner-centred, mixed definition. - Not necessarily being there - The best videos captures the participant with or without instructor in focus, BBC documentaries are some of the best format. The function of video is not to feature talking heads but rather to illustrate and convey complex topic in a way that captures learner's attention. - Just my opinion, but instructor-centeredness is not really a feature that can 'catch learners' attention' but rather a way to personalize and give a 'human' to what is essentially only a human-computer interaction. so the instructors are there, their videos are there to link the students to the human behind all the text and graphs and what not, and that is the bigger function than just 'catching attention'

Based on these comments, the *Instructor-Centred* factor was removed from the category of *Video Quality*. The survey also put an open ended question on every categories that the respondent may proposed other factors that contribute to the quality of MOOC web content. The factors were accepted and justified based on its relevancy and suitability as shown on Table IV. After rigorous study, only one proposed factor is justified based on its relevancy to be considered as one of a web content quality factor for MOOC. The revised initial quality model now consisted of 9 categories and 52 factors when *Instructor-Centred* have been removed, besides *Sound Clarity* and *Light* have been added.

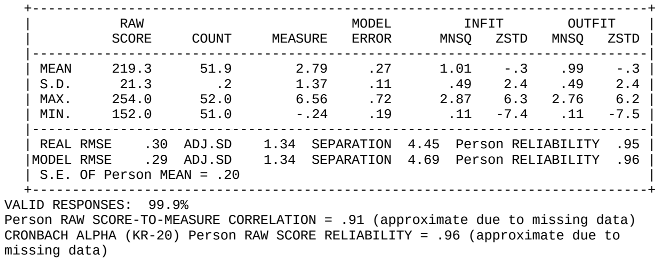
B. Survey on Acceptability

The survey on acceptability was conducted to measure the level of acceptability among content providers and experts towards the content-validated quality model. The summary statistics in Fig. 6 depicts the summary statistics of 47 responses to the 52 web content quality factors by person. The person's mean of +2.79 (SE .27) indicates that majority of respondents found this questionnaire relatively understandable, while showing that their selection was made correctly. This also means that they tend to accept all the proposed factors. The valid responses of 99.9% indicate almost all of the selected respondents are reliable and understand the field with no extreme value. The person reliability (Rasch equivalent to Cronbach's Alpha) is 0.96, indicates high internal consistency of response, which the same result can be expected when the same test is performed.

Item reliability of 0.82 indicates the adequacy of the item to measure what needs to be measured as shown on Fig. 7. The high quality of the items resulted a large value of person separation (4.69) which evidenced by this summary. That's mean that it able to separate person classification that choose "Strongly Agree" to "Strongly Disagree", which provide evidence for external aspects of construct validity as explained by [21]. The mean square fit (Infit and Outfit MNSQ) and the *z* statistics (Infit ZSTD and Outfit ZSTD) for items and persons are closer to their expected values, +1 and 0, respectively. This shows a satisfactory fit to the model.

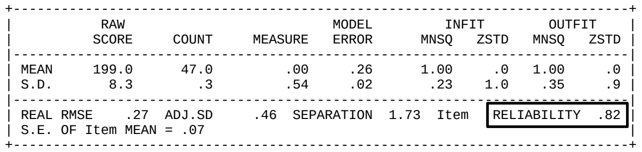
TABLE. IV. PROPOSED QUALITY FACTORS BY RESPONDENT

Category	**Proposed Factor**	**Justification**
Massiveness	Fairness – The content must be equally delivered.	The *Personalized* factor in *Massiveness* category meant to provide relevant information based on learner's personal data, which is gathered throughout the learning process. This factor is sufficient to ensure the fairness of the content. Hence, this proposed factor is rejected.
Openness	Subtitle – To deaf and slow learners	This feature is being taken care by *accessible* factor in *Massiveness* category, which provides access to learners with different abilities, with not only subtitle but also vision and speech. Hence, this proposed factor is rejected.
Video Quality	Recorded -Original video should be recorded, not only use existing one from YouTube	There are *Original* factor in *Conceptual* category which content meant to be developed by the authentic instructor or developer without alteration, deletion or corruptions by any parties. Hence, this proposed factor is rejected.
	Presenter Information - Info of the presenter need to be displayed, not only audio and slides to promote their expertise	This feature is similar with *Relevance* in *Conceptual* category, which stated the content's objective, information and outcome is clear and relevance to the syllabus, learner's requirement and level of study. Hence, this proposed factor is rejected.
	Sound Clarity – The video must have clear sound	*Sound quality* have been highlighted by [19] as one of the successful factor of MOOC web content. Hence, this proposed factor is accepted.
Usability	Report Analysis - Able to recall report certain segments etc.	There is a quality factor named *Analyzable* in *Engagement* category. Hence, this proposed factor is rejected.
Maintaina-bility	Light - Does not consume a lot of resources for mobile and computers	There is a quality factor named *Segmented* in *Video Quality* category. Since lightweight features has been much highlighted such as [20], this factor is accepted and added to *Maintainability.*



```
+------------------------------------------------------------------------------+
|          RAW                              MODEL        INFIT        OUTFIT   |
|         SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|------------------------------------------------------------------------------|
| MEAN     219.3      51.9        2.79     .27      1.01    -.3    .99    -.3  |
| S.D.      21.3        .2        1.37     .11       .49    2.4    .49    2.4  |
| MAX.     254.0      52.0        6.56     .72      2.87    6.3   2.76    6.2  |
| MIN.     152.0      51.0        -.24     .19       .11   -7.4    .11   -7.5  |
|------------------------------------------------------------------------------|
| REAL RMSE    .30  ADJ.SD    1.34  SEPARATION  4.45  Person RELIABILITY  .95  |
|MODEL RMSE    .29  ADJ.SD    1.34  SEPARATION  4.69  Person RELIABILITY  .96  |
| S.E. OF Person MEAN = .20                                                    |
+------------------------------------------------------------------------------+
VALID RESPONSES:  99.9%
Person RAW SCORE-TO-MEASURE CORRELATION = .91 (approximate due to missing data)
CRONBACH ALPHA (KR-20) Person RAW SCORE RELIABILITY = .96 (approximate due to
missing data)
```

Fig. 6. Summary Statistics of the Survey on Acceptability by Person.



```
+------------------------------------------------------------------------------+
|          RAW                              MODEL        INFIT        OUTFIT   |
|         SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|------------------------------------------------------------------------------|
| MEAN     199.0      47.0         .00     .26      1.00     .0   1.00     .0  |
| S.D.       8.3        .3         .54     .02       .23    1.0    .35     .9  |
|------------------------------------------------------------------------------|
| REAL RMSE    .27  ADJ.SD     .46  SEPARATION  1.73  Item  RELIABILITY  .82   |
| S.E. OF Item MEAN = .07                                                      |
+------------------------------------------------------------------------------+
```

Fig. 7. Summary Statistics of the Survey on Acceptability by Items.

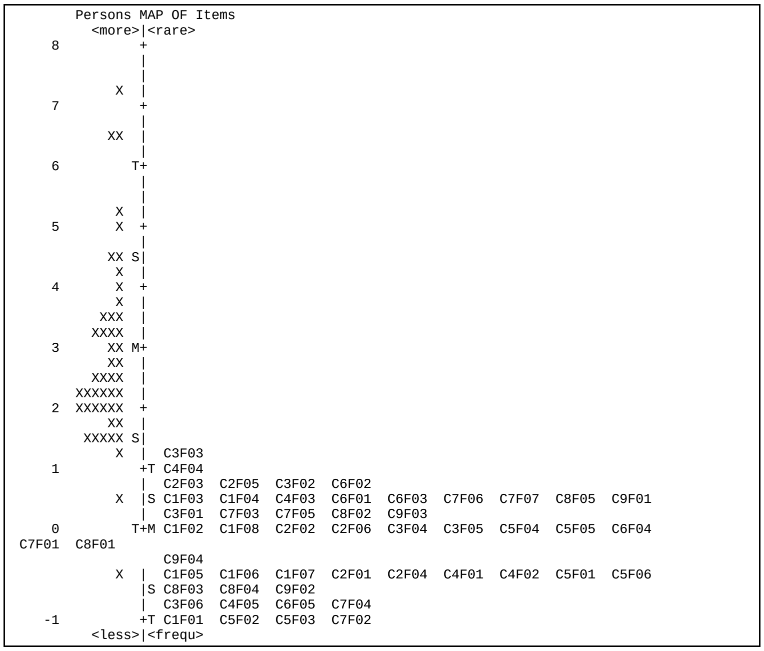
Fig. 8 shows the Wright Map representation for the survey on acceptability of web content quality factors for MOOC. The map shows the distribution that consist the respondent on the left and the item agreement on the right. This map shows that the item mean is significantly below the person mean. In fact, almost all items are located below all persons. This substantially indicates that the majority of respondents understands and tends to agree with the items or factors proposed.

The Wright Map item positioning is simplified by Item Measure Table demonstrated in Fig. 9. The table lists all logits measurement information for each item including mean square (MNSQ), ZSTD value and Point Measure Correlation (PMC). Aligned with the Wright Map, the easiest item to be accepted is at the bottom, which is C5F03 (*Understandable*) while the most difficult item to be accepted is on the top, which is C3F03 (*Translatable*). Both items located respectively at -1.05 and +1.13. The fit statistics of the item is evaluated by MNSQ, which theoretically indicate the accuracy and predictability of the data. The expected value for MNSQ is 1.0 where any values less than 1.0 indicate the observations are too predictable, while greater than 1.0 indicate unpredictability.

According to [16], the acceptable range for Infit and Outfit MNSQ to be considered productive for measurement is between 0.4 to 1.6. On the other hand, the acceptable range for Infit and Outfit ZSTD is between -2.0 to 2.0. According to the scale, three items were identified as misfits namely C8F04 *(Backup ready)* for Infit along with C2F01 *(Multi-Platform)* and C4F01 *(Segmented)* for Outfit. All the misfits also caused the ZSTD value to fall out of reasonable predictability range. Point-correlation is perfect as every item's PMC value is greater than zero, which indicates that all response-level scoring are makes sense.

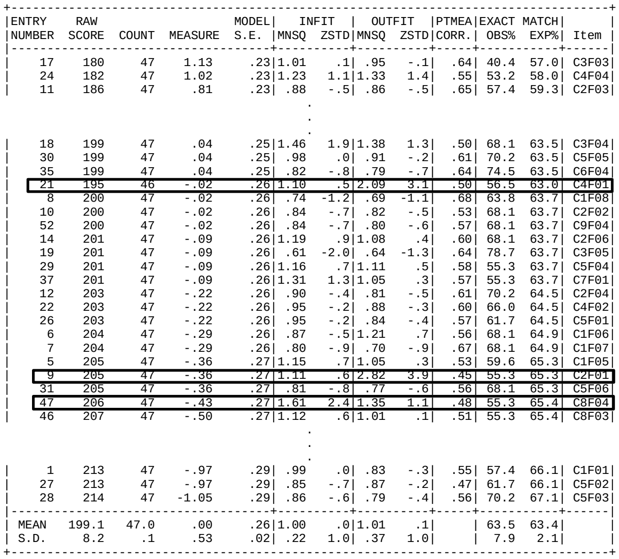
The reevaluation of the three misfit items started with C8F04 (*Backup ready*). The Infit MNSQ rating for this item is 1.61, the value that clearly over the range of productive for measurement, which is 1.6. Therefore, there is high probability that some agreeable person was careless in responding the item. This prediction is strengthen with its high ZSTD value, which is +2.4. The other two misfit items, namely C2F01 (*Multi-Platform*) and C4F01 (*Segmented*) which indicates by overly outfit value may be due to imputed response, lucky guess or careless mistakes. The Guttman Scalogram was referred to detail the misfits. Reference [16] suggests that any suspected responses can be replaced with a missing or blank values before examining the impact of changed result on measures. The crosschecking process on Guttman Scalogram showed that C8F04 (Item 47) was overrated by person A16, while C4F01 (Item 21) and C2F01 (Item 9) was overrated by person A37 and A28 respectively. Therefore, all suspected responses in the dataset were replaced with missing values as suggested.

After performing the suspected responses replacement process, the dataset was retested and the result is illustrated in Fig. 10. Items that were classified as misfit in the first test became fit to the model without distorting the results of other items. For instance, the Infit MNSQ value of C8F04 (*Backup ready*) was adjusted from 1.61 to 1.48, resulting the decrement of ZSTD value from 2.4 to 2.0 to put it within the reasonable predictability range. The C4F01 (*Segmented*) and C2F01 (*Multi-Platform*) values of MNSQ were moved to the acceptable range due to the replacement process. Contrarily, the ZSTD value of item C4F03 is still over the acceptable range (-2.0 to 2.0) which is 2.2. However, as the value of Infit MNSQ is within range, the item was validated.



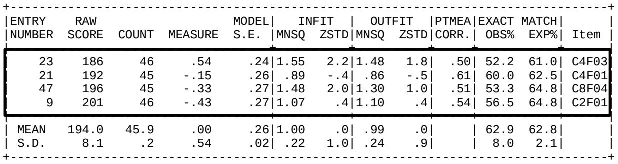
```
       Persons MAP OF Items
          <more>|<rare>
    8           +
                |
                |
           X    |
    7           +
                |
          XX    |
                |
    6          T+
                |
                |
           X    |
    5      X    +
                |
          XX   S|
           X    |
    4      X    +
           X    |
         XXX    |
        XXXX    |
    3     XX   M+
          XX    |
        XXXX    |
      XXXXXX    |
    2 XXXXXX    +
          XX    |
       XXXXX   S|
           X    |  C3F03
    1           +T C4F04
                |  C2F03  C2F05  C3F02  C6F02
           X    |S C1F03  C1F04  C4F03  C6F01  C6F03  C7F06  C7F07  C8F05  C9F01
                |  C3F01  C7F03  C7F05  C8F02  C9F03
    0          T+M C1F02  C1F08  C2F02  C2F06  C3F04  C3F05  C5F04  C5F05  C6F04
C7F01  C8F01
                   C9F04
           X    |  C1F05  C1F06  C1F07  C2F01  C2F04  C4F01  C4F02  C5F01  C5F06
                |S C8F03  C8F04  C9F02
                |  C3F06  C4F05  C6F05  C7F04
   -1           +T C1F01  C5F02  C5F03  C7F02
          <less>|<frequ>
```

Fig. 8. The Wright Map of the Survey on Acceptability by Items.



```
+-----------------------------------------------------------------------------+
|ENTRY    RAW                          MODEL|   INFIT  |  OUTFIT  |PTMEA|EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.| OBS%  EXP%| Item |
|------------------------------------+----------+----------+-----+-----------+------|
|    17    180     47     1.13      .23|1.01     .1| .95    -.1|  .64| 40.4  57.0| C3F03|
|    24    182     47     1.02      .23|1.23    1.1|1.33    1.4|  .55| 53.2  58.0| C4F04|
|    11    186     47      .81      .23| .88    -.5| .86    -.5|  .65| 57.4  59.3| C2F03|
                                      .
                                      .
                                      .
|    18    199     47      .04      .25|1.46    1.9|1.38    1.3|  .50| 68.1  63.5| C3F04|
|    30    199     47      .04      .25| .98     .0| .91    -.2|  .61| 70.2  63.5| C5F05|
|    35    199     47      .04      .25| .82    -.8| .79    -.7|  .64| 74.5  63.5| C6F04|
|    21    195     46     -.02      .26|1.10     .5|2.09    3.1|  .50| 56.5  63.0| C4F01|
|     8    200     47     -.02      .26| .74   -1.2| .69   -1.1|  .68| 63.8  63.7| C1F08|
|    10    200     47     -.02      .26| .84    -.7| .82    -.5|  .53| 68.1  63.7| C2F02|
|    52    200     47     -.02      .26| .84    -.7| .80    -.6|  .57| 68.1  63.7| C9F04|
|    14    201     47     -.09      .26|1.19     .9|1.08     .4|  .60| 68.1  63.7| C2F06|
|    19    201     47     -.09      .26| .61   -2.0| .64   -1.3|  .64| 78.7  63.7| C3F05|
|    29    201     47     -.09      .26|1.16     .7|1.11     .5|  .58| 55.3  63.7| C5F04|
|    37    201     47     -.09      .26|1.31    1.3|1.05     .3|  .57| 55.3  63.7| C7F01|
|    12    203     47     -.22      .26| .90    -.4| .81    -.5|  .61| 70.2  64.5| C2F04|
|    22    203     47     -.22      .26| .95    -.2| .88    -.3|  .60| 66.0  64.5| C4F02|
|    26    203     47     -.22      .26| .95    -.2| .84    -.4|  .57| 61.7  64.5| C5F01|
|     6    204     47     -.29      .26| .87    -.5|1.21     .7|  .56| 68.1  64.9| C1F06|
|     7    204     47     -.29      .26| .80    -.9| .70    -.9|  .67| 68.1  64.9| C1F07|
|     5    205     47     -.36      .27|1.15     .7|1.05     .3|  .53| 59.6  65.3| C1F05|
|     9    205     47     -.36      .27|1.11     .6|2.82    3.9|  .45| 55.3  65.3| C2F01|
|    31    205     47     -.36      .27| .81    -.8| .77    -.6|  .56| 68.1  65.3| C5F06|
|    47    206     47     -.43      .27|1.61    2.4|1.35    1.1|  .48| 55.3  65.4| C8F04|
|    46    207     47     -.50      .27|1.12     .6|1.01     .1|  .51| 55.3  65.4| C8F03|
                                      .
                                      .
                                      .
|     1    213     47     -.97      .29| .99     .0| .83    -.3|  .55| 57.4  66.1| C1F01|
|    27    213     47     -.97      .29| .85    -.7| .87    -.2|  .47| 61.7  66.1| C5F02|
|    28    214     47    -1.05      .29| .86    -.6| .79    -.4|  .56| 70.2  67.1| C5F03|
|------------------------------------+----------+----------+-----+-----------+------|
| MEAN   199.1   47.0      .00      .26|1.00     .0|1.01     .1|     | 63.5  63.4|      |
| S.D.     8.2     .1      .53      .02| .22    1.0| .37    1.0|     |  7.9   2.1|      |
+-----------------------------------------------------------------------------+
```

Fig. 9. Part of Item Measure for Survey on Acceptability with Highlighted Misfit Items.



```
+-----------------------------------------------------------------------------+
|ENTRY    RAW                          MODEL|   INFIT  |  OUTFIT  |PTMEA|EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.| OBS%  EXP%| Item |
|------------------------------------+----------+----------+-----+-----------+------|
|    23    186     46      .54      .24|1.55    2.2|1.48    1.8|  .50| 52.2  61.0| C4F03|
|    21    192     45     -.15      .26| .89    -.4| .86    -.5|  .61| 60.0  62.5| C4F01|
|    47    196     45     -.33      .27|1.48    2.0|1.30    1.0|  .51| 53.3  64.8| C8F04|
|     9    201     46     -.43      .27|1.07     .4|1.10     .4|  .54| 56.5  64.8| C2F01|
|------------------------------------+----------+----------+-----+-----------+------|
| MEAN   194.0   45.9      .00      .26|1.00     .0| .99     .0|     | 62.9  62.8|      |
| S.D.     8.1     .2      .54      .02| .22    1.0| .24     .9|     |  8.0   2.1|      |
+-----------------------------------------------------------------------------+
```

Fig. 10. Item Measure after Suspected Response Replacement Process.

There are open-ended questions in the survey about the other factor needed to determine quality of MOOC web content, but no significant comment was provided by the respondents. The probability of factors to be accepted by the respondents had been calculated based on the logits value of Item Measure. The result shown on Table V clearly proved that the probability of all factors to be accepted by respondents on average was exceeds 70% of Cronbach's Alpha. This means that all factors are significantly acceptable to determine the web content quality for MOOC.

Rasch analysis also utilized to determine the validity of the used scale by making a zero setting and calibrating the rating scale as presented on Fig. 11. Besides, it determines that the probability of response distribution is equal between the specified scales (equal interval). The increases in value of *observe average*, indicates normal response pattern as depicted on Fig. 12. *Structure Calibration* in turn solves the problem of elasticity of gaps within the Likert scale threshold. In this analysis, it has been proved that all deviation values are within the range $1.4<s<5.0$. The calculation is as follows:

$s_{1\text{-}2}$: 0.00 – 2.97 = 2.97 > 1.4

$s_{2\text{-}3}$: 2.97 – 0.84 = 2.13 > 1.4

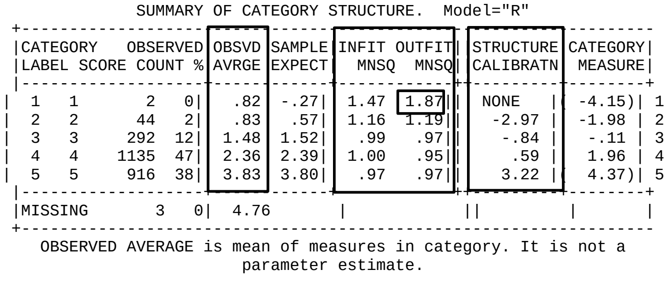
$s_{3\text{-}4}$: 0.84 – (-0.59) = 1.43 > 1.4

$s_{4\text{-}5}$: 3.22 – 0.59 = 2.63 > 1.4

Fig. 11 also shows that the person and item data fitness were also manageable as the Infit and Outfit MNSQ is all in the productive range, except for scale 1 (Outfit MNSQ 1.87). However, it's also validated since the value is not degrading as agreed by [16].

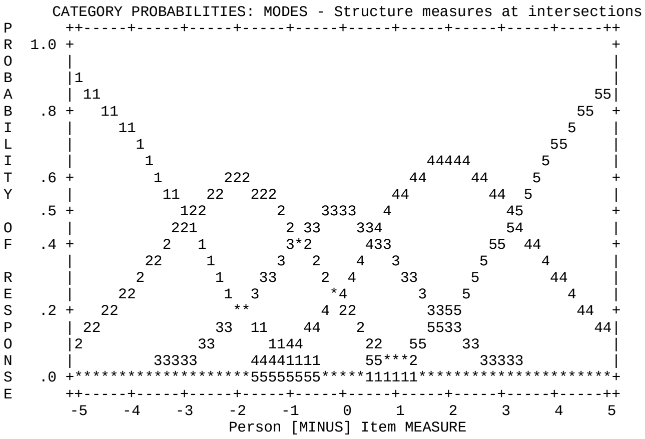
TABLE. V. THE PROBABILITY OF FACTORS TO BE ACCEPTED BY RANKING (TOP 10 FACTORS)

Code	Factor Name	Person Measure	Item Measure	P(Ө) %
C1F01	Relevance	2.88	-0.97	97.92
C5F02	Readable	2.88	-0.97	97.92
C7F02	Available	2.88	-0.89	97.75
C3F06	Flexible	2.88	-0.81	97.56
C4F05	Clear Audio	2.88	-0.81	97.56
C6F05	Visible	2.88	-0.81	97.56
C7F04	Reliable	2.88	-0.65	97.15
C9F02	Continuous Improvement	2.88	-0.58	96.95
C8F03	Secure	2.88	-0.5	96.71
C2F01	Multi-Platform	2.88	-0.43	96.48



```
                  SUMMARY OF CATEGORY STRUCTURE.  Model="R"
 +------------------------------------------------------------------
 |CATEGORY   OBSERVED|OBSVD SAMPLE|INFIT OUTFIT||STRUCTURE|CATEGORY|
 |LABEL SCORE COUNT %|AVRGE EXPECT|  MNSQ  MNSQ||CALIBRATN| MEASURE|
 |-------------------+------------+------------++---------+--------+
|  1   1       2   0|   .82  -.27|  1.47  1.87||  NONE   |( -4.15)| 1
|  2   2      44   2|   .83   .57|  1.16  1.19||   -2.97 | -1.98  | 2
|  3   3     292  12|  1.48  1.52|   .99   .97||    -.84 |  -.11  | 3
|  4   4    1135  47|  2.36  2.39|  1.00   .95||     .59 |  1.96  | 4
|  5   5     916  38|  3.83  3.80|   .97   .97||    3.22 |(  4.37)| 5
 |-------------------+------------+------------++---------+--------+
 |MISSING       3   0|  4.76      |            ||         |        |
 +------------------------------------------------------------------
    OBSERVED AVERAGE is mean of measures in category. It is not a
                        parameter estimate.
```

Fig. 11. Rating Scale (Partial Credit).



```
      CATEGORY PROBABILITIES: MODES - Structure measures at intersections
P      ++-----+-----+-----+-----+-----+-----+-----+-----+-----+-----++
R  1.0 +                                                             +
O      |                                                             |
B      |1                                                            |
A      | 11                                                        55|
B   .8 +   11                                                    55  +
I      |     11                                                 5    |
L      |       1                                              55     |
I      |        1                               44444        5       |
T   .6 +         1       222                  44     44     5        +
Y      |          11   22   222             44          44  5        |
    .5 +            122        2    3333   4             45          +
O      |           221          2 33    334              54          |
F   .4 +          2   1         3*2      433           55  44        +
       |        22     1       3   2    4   3         5      4       |
R      |       2        1    33     2  4     33       5        44     |
E      |     22          1  3        *4        3    5           4    |
S   .2 +   22             **        4 22        3355             44  +
P      | 22             33  11    44    2       5533               44|
O      |2             33      1144       22   55    33               |
N      |         33333      44441111     55***2       33333          |
S   .0 +********************55555555*****111111**********************+
E      ++-----+-----+-----+-----+-----+-----+-----+-----+-----+-----++
        -5    -4    -3    -2    -1    0     1     2     3     4     5
                          Person [MINUS] Item MEASURE
```

Fig. 12. Category Probabilities Modes.

Rasch set the minimum value of *raw variance explained by measure* to 40% to be accepted as a benchmark to ensure unidimensionality in this model [22]. As shown in Fig. 13, the model's *raw variance explained by measure* value is 60.7%, indicates that it has good unidimensionality feature. The value of *Unexplained variance in 1st contrast* indicates that there is a bit disruption to the items, known as noise. However, the percentage is very low at 4.9%, compared to the maximum controlled value of 15% as pointed out by [23]. This is confirmed by the table of largest standardized residual correlations as shown in Fig. 14. The table indicates that there is no locally dependant pairs of items which having residual correlation > .7, as the largest residual correlation is only .53.

As a discussion, the proposed model validation has been executed through content validity test and the survey on acceptability. The Rasch Model was utilized to prove two things (i) Data fitness (ii) The probability of the quality factors to be accepted. The data fitness is proven by statistical analysis on infit / outfit MNSQ and ZSTD, which is all in the productive range to be measured. The Wright Map and Item Measure Table not only assists the data fitness analysis but also the level of agreement determination for every item, by placing the most agreed item at below and least agreed item on above. This enables rearrangement of the factors for each categories in the quality model according to the level of agreement as indicated by survey on acceptability. Every factors definition was also revised based on the result of model validation processes. The final web content quality model for MOOC was devised as depicted in Fig. 15.

Besides, several Rasch Model features such as Category Probability Modes and Principal Component Analysis assist the determination of item unidimensionality, which means that all items the questionnaire measure only a single construct. The feature is critical especially in forming a newly-developed hierarchical model, like the one we developed and validated in this research. The result of Principal Component Analysis prove that the model developed in this research is completely hierarchical with each criterion related to only one family, similar with other hierarchical models like ISO/IEC 9126.

C. Threats to Validity

There are several issues that may threatening the validity of the result and model. Thus, four types of threats to the validity of the survey were analysed based on framework proposed by [24] which is internal, external, conclusion and construct. The narrowly focused purposive sampling utilized for this study strengthens the trustworthy inference, which increases the internal validity. The selection of respondents was also carefully undertook and reconsidered by the field experts before the content validity test and survey on acceptability were carried out.

Threat to external validity are manageable as the value of item and person reliability in content validity test and survey on acceptability is beyond the standard of Cronbach Alpha which is 0.7 [25]. The reliability score of 0.95 for person in the survey of acceptability indicates the consistency of the result and generalizable outside the respondents setting. In term of conclusion validity, the measurement used to analyze data is considered reliable by the application of the Rasch Model. Moreover, the high reliability score for item which is 0.82 proves data sufficiency to measure what should be measured, thus guaranteeing the conclusion validity. Threats to construct validity are taken care by utilization of Rasch Measurement Model to prove unidimensionality feature of the survey result as well as the proposed model. The evidence is when the value of raw variance explained by measure value beyond Rasch model of 60% which is 80.2%. The items that fit are likely to be measuring the single dimension intended by the construct theory.

```
             TABLE 23.3 Survey Acceptability Test2 190819
                  ZOU661WS.TXT Sep 8  0:49 2019
   INPUT: 47 Persons 52 Items  MEASURED: 47 Persons  52 Items  5 CATS
                              1.0.0
-------------------------------------------------------------------------

           CONTRAST 1 FROM PRINCIPAL COMPONENT ANALYSIS OF
    STANDARDIZED RESIDUAL CORRELATIONS FOR Items (SORTED BY LOADING)
     Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                          Empirical       Modeled
Total variance in observations     =        132.2 100.0%         100.0%
Variance explained by measures     =         80.2  60.7%          60.0%
Unexplained variance (total)       =         52.0  39.3% 100.0%   40.0%
    Unexplned variance in 1st contrast =          4.9   3.7%   9.4%
```

Fig. 13. Principal Component Analysis.

```
+----------------------------------+
|RESIDUL| ENTRY        | ENTRY       |
|CORRELN|NUMBER Item   |NUMBER Item  |
|-------+-------------+-------------|
|   .53 |    14 C2F06 |    31 C5F06 |
|   .53 |    18 C3F04 |    19 C3F05 |
|   .51 |    16 C3F02 |    49 C9F01 |
|   .50 |    11 C2F03 |    14 C2F06 |
|   .47 |    21 C4F01 |    22 C4F02 |
|   .44 |    38 C7F02 |    40 C7F04 |
|   .43 |     9 C2F01 |    19 C3F05 |
|-------+-------------+-------------|
|  -.47 |    11 C2F03 |    27 C5F02 |
|  -.46 |    36 C6F05 |    52 C9F04 |
|  -.46 |    29 C5F04 |    39 C7F03 |
+----------------------------------+
```

Fig. 14. Largest Standardized Residual Correlation.

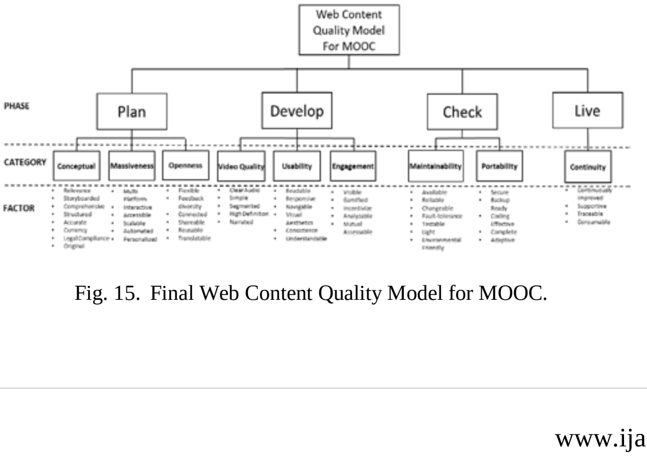


Fig. 15. Final Web Content Quality Model for MOOC.

V. Conclusion

This research demonstrates the effectiveness of two validation techniques which are: (1) content validity test and (2) survey on acceptability to verify the data fitness and probability of acceptance for the web content quality model. The content validity test was used to confirm whether the content of the survey is acceptable to the reviewers, which provides empirical evidence to the construct validity. A proposed factor which is *Instructor-centred* was excluded, while two new factors were proposed by the respondents, which is *Sound Quality* and *Light*. Then, the survey on acceptability was conducted to measure the probability of acceptance of every category and factor for the quality model based on the perspective of content providers and experts.

In order to provide evidence to construct validity, Rasch Model was applied to provide hypothetical unidimensional line along items and persons according to their difficulty and ability. The Rasch application built-in tools like the Wright Map and the Guttman Scalogram facilitate the determination of data fitness and probability of acceptance for every item which being measured in intervals via logits. While this approach claimed to be revolutionary in statistical application, this research proves it suitability for construct validation and instrument development for the development of a quality model. Besides, the features like Category Probability Modes and Principal Component Analysis assist the determination of item unidimensionality, which means that all items measure only a single construct, the feature which very pertinent in developing a new hierarchical model.

Acknowledgment

The research is sponsored by Malaysia Ministry of Higher Education under Fundamental Research Grant Scheme (FRGS) and the program of Federal Training Award (HLP).

References

[1] Sunar, S. White, N. Abdullah, and H. Davis, “How learners’ interactions sustain engagement: a MOOC case study,” IEEE Trans. Learn. Technol., vol. X, no. X, pp. 1–1, 2016.

[2] W. N. W. A. Rahman, H. Zulzalil, I. Ishak, and A. W. Selamat, “Quality Model for Massive Open Online Course (MOOC) Web Content,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 10, no. 1, 2020.

[3] S. Messick, “Validity of Psychological Assessment: Validation of Inferences from Persons’ Responses and Performances as Scientific Inquiry into Score Meaning,” Am. Psychol., vol. 50, no. 9, 1995.

[4] A. A. Aziz, A. Mohamed, and N. Arshad, “Application of Rasch Model in validating the construct of measurement instrument,” vol. 2, no. 2, 2008.

[5] Z. M. Zain, A. A. A. Ghani, R. Abdullah, and R. Atan, “Blog Quality Model,” Int. J. Web Based Communities, vol. 9, no. 1, pp. 25–50, 2013.

[6] F. De Battisti, G. Nicolini, and S. Salini, “The Rasch Model in Customer Satisfaction Survey Data,” vol. 3703, 2016.

[7] T. G. Bond and C. M. Fox, “Applying the Rasch Model : Fundamental Measurement in the Human Sciences Second Edition University of Toledo,” 2007.

[8] A. Caro, C. Calero, I. Caballero, and M. Piattini, “A proposal for a set of attributes relevant for Web portal data quality,” Softw. Qual. J., pp. 513–542, 2008.

[9] A. A. Shah, S. D. Ravana, S. Hamid, and M. A. Ismail, “Web credibility assessment : affecting factors and assessment techniques,” Inf. Res. J., vol. 20, pp. 1–28, 2015.

[10] E. T. Loiacono, R. T. Watson, and D. L. Goodhue, “WebQual TM : A Measure of Web Site Quality,” Mark. theory Appl., vol. 13, no. 3, pp. 432–438, 2002.

[11] M. Abdellatief, A. B. Sultan, M. A. Jabar, and R. Abdullah, “A Technique for Quality Evaluation of E-Learning from Developers Perspective A Technique for Quality Evaluation of E-Learning from Developers Perspective,” no. May 2018, 2011.

[12] J. Shi, X. Mo, and Z. Sun, “Content validity index in scale development,” J. Cent. South Univ. Med. Sci., vol. 37, no. 2, p. 152—155, Feb. 2012.

[13] B. A. Kitchenham and S. L. Pfleeger, “Personal Opinion Surveys,” Guid. to Adv. Empir. Softw. Eng., pp. 63–92, 2008.

[14] A. Khanjani, “Quality of Service Model for Software as a Service in Cloud Computing from users’ and Providers’ Perspectives,” UPM, 2015.

[15] W. J. Boone, “Rasch analysis for instrument development: Why,when,and how?,” CBE Life Sci. Educ., vol. 15, no. 4, 2016.

[16] J. M. Linacre, “What do Infit and Outfit, Mean-square and Standardized mean?,” Rasch Measurement Transactions, 2002. [Online]. Available: https://www.rasch.org/rmt/rmt162f.htm.

[17] S. M. Sincero, “Online Surveys,” Explorable.com, 2012. [Online]. Available: https://explorable.com/online-surveys. [Accessed: 11-Dec-2019].

[18] N. V. Raju and N. S. Harinarayana, “Online survey tools : A case study of Google Forms Online Survey Tools : A Case Study of Google Forms,” in Scientific, Computational & Information Research Trends in Engineering, GSSS-IETW, Mysore, 2018, no. December.

[19] A. M. F. Yousef, M. A. Chatti, U. Schroeder, and M. Wosnitza, “What drives a successful MOOC? An empirical examination of criteria to assure design quality of MOOCs,” Proc. - IEEE 14th Int. Conf. Adv. Learn. Technol. ICALT 2014, pp. 44–48, 2014.

[20] S. I. De Freitas, J. Morgan, and D. Gibson, “Will MOOCs transform learning and teaching in higher education? Engagement and course retention in online learning provision,” Br. J. Educ. Technol., vol. 46, no. 3, pp. 455–471, 2015.

[21] Z. M. Zain, A. Azim, A. Ghani, R. Abdullah, and R. Atan, “Blog Quality Measurement : Analysis of Criteria using The Rasch Model,” vol. 1, no. 3, pp. 665–682, 2011.

[22] W. P. Fisher, “Rating scale instrument quality criteria,” Rasch Meas. Trans., vol. 21, no. 1, p. 1095, 2007.

[23] A. A. Aziz, Mohd Saidfudin Masodi, and Azami Zaharim, Basic of Rasch Measurement Model. UKM Publisher, 2013.

[24] C. Wohlin, P. Runeson, M. H¨ost, M. C. Ohlsson, B. Regnell, and A. Wessl´en, Experimentation in Software Engineering. 2000.

[25] K. S. Taber, “The Use of Cronbach ’ s Alpha When Developing and Reporting Research Instruments in Science Education,” 2016.

Prediction Intervals based on Doubly Type-II Censored Data from Gompertz Distribution in the Presence of Outliers

S. F. Niazi Alil[1]
Mathematics Department,
Faculty of Science and Human Studies
of Hotat Sudair, Majmaah University,
Majmaah 11952, Saudi Arabia.
Mathematics Department,
Faculty of Science, Al-Azhar University,
Assuit branch, 71524 Assiut, Egypt.

Ayed R. A. Alanzi[2]
Mathematics Department,
Faculty of Science and Human Studies
of Hotat Sudair, Majmaah University,
Majmaah 11952, Saudi Arabia.

***Abstract*—The study aims at getting the Bayesian predication intervals for some order statistics of future observations from the distribution of Gompertz (Gomp (α, β)). Doubly Type-II censored data has assisted obtaining in the presence of single outlier that arose from the different same family members of distribution. Single outlier of type $\beta\,\beta_0$ and $\beta+\beta_0$ are considered and bivariate independent prior density for α and β are used. The problem of solving the Double integral to obtain the closed form for α and β, leads us to use MCMC for calculating the Bayesian Predication Intervals. The use of numerical examples and statistical data has enable to properly present and describe the procedure. We conclude that the Bayesian predication intervals are shorter for y_1 than y_5 when we are increasing the β_0 value.**

Keywords*—Bayesian prediction; Gompertz distribution; predictive distribution; doubly Type-II censored data; Markov Chain Monte Carlo; single outliers

I. Introduction

The adult death patterns can be effectively described through the use of the Gompertz distribution ([17]; [6]). Moreover, the Gompertz mortality force for the decreased infant and young adult levels of mortality extends to the whole life population span without any observed deceleration of mortality ([16]). A continuous probability density function (pdf) and a cumulative distribution function (cdf) are the constituents of the Gompertz distribution.
The pdf as follows:

$$f(x) = \alpha\,\beta\,e^{\alpha\,x - \beta\,(e^{\alpha\,x} - 1)}, \quad x > 0, \alpha > 0, \beta > 0, \qquad (1)$$

and The cdf as follows:

$$F(x) = 1 - e^{-\beta\,(e^{\alpha\,x} - 1)}. \qquad (2)$$

This distribution should be denoted with two Gomp α and β parameters. The research conducted by [1] indicated that a simple transformation relates the Gompertz distribution to a certain distribution in the family of distributions. A further research conducted by [7] showed that it is possible to get the maximum likelihood parameter estimates the Gompertz model. The study by [3] suggests the ways to apply it and provides a more recent survey that enables to better understand the model. At the same time, [19] made an attempt to reformulate the Gompertz mortality force and get an insight into the new formation relationship.

The analysis of the research by [18] enabled to trace the connections between the Weibull, the Gompertz, and other Type I extreme value distributions. Later, [9] managed to obtain a Bayesian prediction, mixing two-component lifetime model of Gompertz. In another study [10] derived a Bayesian record statistics analysis from the Gompertz model. A negative Gompertz distribution was presented by later, [11] who focused on the discussion of the negative aging parameter rate. A generalized three-parameter Gompertz distribution was presented by [8]), who provided a deep insight into the topic under investigation. Furthermore, [2]worked on the Gompertz model, and attempted to introduce a more generalized four-parameter version of the model that was referred to as a beta-Gompertz distribution. Also, the paper provides some commonly used distributions, including generalized and beta-exponential Gompertz distributions as sub-models. [15] proposed a distribution of an exponentiated Weibull extension; however, it was modified. It was further generalized and discussed in the study by [8]. Author in [13] focused on the investigation and discussion of the obtained prediction intervals that are based on Gompertz doubly censored data. There are some cases make Progressive Hybrid Censored schemes (PHCS) difficult to apply when the failures may occur before time [21]. Some researchers estimated and predicted the Generalized Progressive Hybrid Censored Data for Gompertz Distribution [20]. Whoever Gompertz distribution was studied by many researchers such as [22].

The main objective of this paper, we assume that $X_I, X_2, \cdots, X_n$, is an ordered random sample of size n drawn from a population whose pdf, is Gomp(α, β), which is defined by equation 1, and that $Y_1, Y_2, \cdots, Y_m$. is a second independent random sample (of size m) of future observations from the same distributions. Bayesian prediction bounds for the future observations $Y_t, Y_z, \cdots, Y_m$ in the presence of a single outlier of type $\beta\,\beta_0$ and $\beta + \beta_0$ are obtained.

Observation is an outlier in the data set that is inconsistent with the data set remainder ([5]). Hence, a single $\beta\,\beta_0$, and $\beta + \beta_0$ type outliers are present in the future Gompertz population sample. Gomp$(\alpha,\, \beta\,\beta_0)$ is taken for a single type $\beta\, +\, \beta_0$ outlier of the pdf, while in the case of single type $\beta + \beta_0$ outlier the pdf is taken Gomp$(\alpha,\, \beta\, +\, \beta_0)$.

In the study, the bounds of the Bayesian prediction are received for the future Gopm $(\alpha,\, \beta)$ distribution observations in the presence of a single outlier of type. It is considered that both parameters α and β are unknown. The true value (β, α) uncertainty is measured through the function of the bivariate prior density that was discussed and applied with the same model in the research conducted by [10].Furthermore, the current research presupposes the construction of the predictive interval that will be used for the future observation with the presence of a single outlier of type with MCMC. The use of statistics will assist in illustrating and presenting the procedure.

In this article, Section II explains the Likelihood Function. After that Section III discuss the Posterior distribution. Moreover, Section IV clarify the Bayesian predication in the presence of outliers for future observations with two schemes $\beta\,\beta_0$ and $\beta + \beta_0$. Section V shows numerical example, which are consider the previous two schemes. In the final Section VI, we give the conclusion and opens future direction.

II. Likelihood Function

In this section, we assume $x_1,\, x_2,\, \cdots,\, x_n$ is an ordered random size n sample from the Gopm$(\alpha,\, \beta)$. The pdf and cdf are given be (1) and (2) , respectively. Also, let $x_1 \leq x_2 \leq \cdots \leq x_k$ be the k smallest ordered observation, while $x_{r+1} \leq x_{r+2} \leq \cdots \leq x_n$, the $n-r$ largest ordered observations in the sample. The statistical analysis contains the application of only the remaining ordered observations, that is, $\underline{x} = (x_{k+1},\, x_{s+2},\, \cdots,\, x_r)$. Moreover, it is evident that when $k=1$, the sample will be a Type-II right censored sample. A doubly censored sample pulled from population with pdf and cdf as given in (1) and (2) that likelihood function is given as follow:

$$\begin{aligned} L(\alpha, \beta; \underline{x}) \propto\; & [F_X(x_{k+1};\, \alpha,\, \beta)]^k [1 - F_X(x_r;\, \alpha,\, \beta)]^{n-r} \\ & \times \prod_{i=k+1}^{r} [f_X(x_i; \alpha,\, \beta)], x_{s+1}\, \geq 0 \\ =\; & (\alpha\,\beta)^{r-s}[1 - \exp\{-\beta T_1(\alpha;\, x_{k+1})\}]^k \\ & \times \exp\Big\{\alpha \sum_{i=s+1}^{r} x_i - \beta\, T_2(\underline{x}; \alpha)\Big\}. \end{aligned} \tag{3}$$

where

$$\begin{aligned} T_1(\alpha;\, x_{k+1}))\; &= e^{\alpha\, x_{k+1}} - 1, \\ T_2(\alpha;\, \underline{x})\; &= (n-r)e^{\alpha\, x_r} + \sum_{i=k+1}^{r} e^{\alpha\, x_i} - n + s. \end{aligned} \tag{4}$$

The Bayesian prediction tends to bound the future observations in the presence of a single outlier of type $Gomp(\alpha,\, \beta)$ distribution when two parameters types α and β are both dependent and unknown.

III. The Posterior Distribution

To obtain the joint posterior density of α and β, we use a bivariate prior density of the form:

$$\pi(\alpha,\, \beta)\, =\, \pi_1(\alpha)\, \pi_2(\beta), \tag{5}$$

where

$$\pi_1(\alpha)\, =\, \frac{\gamma_1^{\eta_1}}{\Gamma(\eta_1)}\, \alpha^{\eta_1 - 1}\, e^{-\alpha\gamma_1},\, (\eta_1,\, \gamma_1\, >\, 0) \tag{6}$$

and

$$\pi_2(\beta) = \frac{\gamma_2^{\eta_2}}{\Gamma(\eta_2)}\, \beta^{\eta_2 - 1}\, e^{-\beta\gamma_2}\, (\eta_2,\, \gamma_2\, >\, 0). \tag{7}$$

The paper assumes that the joint prior density for the parameter α and β is the form (5) and presented by Jaheen [10] for the progressive censored data prediction from the Gompertz model and applied by [13] for the prediction Gompertz doubly censored data intervals.

The likelihood of the function presented by (3) and the function of the joint prior density presented by (5)as well as the function of the joint posterior density of α and β is

$$\pi^*(\alpha,\, \beta|\underline{x}) \quad = \frac{L(\alpha,\, \beta;\, \underline{x})\pi_1(\alpha)\, \pi_2(\beta)}{\int_0^\infty \int_0^\infty L(\alpha,\, \beta;\, \underline{x})\pi_1(\alpha)\, \pi_2(\beta) d\alpha d\beta}. \tag{8}$$

The joint posterior density function of α and β given data can be written as

$$\pi^*(\beta,\, \alpha, |\underline{x}) \propto h_1(\beta\,|\,\alpha,\, data) h_2(\alpha|\, data) h_3(\alpha, \beta|\, data) \tag{9}$$

where $h_1(\beta\,|\,\alpha,\, data)$ is a gamma density where the shape parameter $m = r - k + \eta_1$ and the scale parameter is $\gamma_1 +\ T_2(\alpha;\, \underline{x})$. At the same time, $h_2(\alpha|\, data)$ is a proper density function of the form

$$h_2(\alpha|data) \propto \qquad \frac{1}{[\gamma_1 + T_2(\alpha;\, \underline{x})]^m}\, \alpha^{r-k+\eta_2-1} \\ e^{-\alpha\,(\frac{1}{\gamma_2} - \sum_{i=k+1}^{r} x_i)} \tag{10}$$

and $h_3(\alpha, \beta|\, data))$ is given by

$$h_3(\alpha,\, \beta|data) = \Big[1 - e^{-\beta T_1(\alpha;\, x_{k+1})}\Big]^s. \tag{11}$$

From equation (8) and it enables to see that a simple closed form cannot express the equation. Therefore, the Bayes estimators of the parameter α and β cannot be received in simple closed forms. Hence, the paper suggests the approximation (9) by applying the importance sampling technique that is also presented by [14]. The importance sampling details are presented below.

In this paper, we used the importance sampling procedure to calculate the Bayes estimates for α, β as well as any function of the parameters $g(\alpha,\, \beta)$. Moreover, the Algorithm 1 (presented below) is used to generate α and β from the posterior density function (7).

Algorithm 1:

Step 1 : Start with an $(\alpha^0;\, \beta^0)$.

Step 2 : set t = 1.

Step 3 : Generate α^t from $h_2(\alpha|\, data)$ using the method developed by [12] with the $N(\alpha^{t-1},\, \sigma)$ proposal distribution, where σ^2 is the variance of the parameter α.

Step 4 : Generate β^t from gamma distribution with pdf $h_2(\beta\,|\, \alpha,\, data)$.

Step 5 : Put t = t+1.

Step 6 : Repeat steps 3-5 M times to obtain $\{(\alpha^t,\, \beta^t),\, t = 1,\, 2,\, \cdots,\, M\}$.

The approximate Bayes are applied to estimate any function of the parameters say $g(\alpha,\, \beta)$ under the squared functions of error loss using the procedure of importance sampling, as shown below:

$$\hat{g}_{BS}(\alpha,\, \beta) \quad = \frac{\sum\limits_{i=1}^{M} g(\alpha_i,\, \beta_i)\, g_3(\alpha_i,\, \beta_i|data)}{\sum_{i=M_0}^{M} g_3(\alpha_i,\, \beta_i|data)}, \tag{12}$$

IV. Bayesian Prediction in the Presence of a Single Outlier for Future Observations

The section introduces the prediction of the future observations in the presence of a single outlier. Also, it is assumed that $X_1,\, X_2,\, \cdots,\, X_n$ is a random size n sample drawn from the $Gomp(\alpha, \beta)$ population, where the pdf is presented by (1). Let us assume that $Y_1,\, Y_2,\, \cdots,\, Y_m$ is a second, independent, unobserved size m sample received from the same population. This sample is the future sample, and the aim of the study is to get Bayesian prediction bounds for the s^{th} oncoming observation Y_s, $s = 1,\, 2,\, \cdots,\, m$ in the presence of a single outlier.

In the case of the size m sample, let Y_s be the s^{th} ordered lifetime, $1 \,\leq\, s \leq\, m$. Then the Y_s density function for a given θ in the presence of a single outlier is of the form $f = f(y|\,\theta)$ and $F = F(y|\theta)$ are the distribution and density functions of all y_s which are not referred to be outliers as $f^* = f^*(y|\theta)$ and $F^* = F^*(y|\theta)$ are those of an outlier ([4]). The f^* and F^* functions are received for the Gomp$(\alpha,\, \beta)$ model through the replacement of parameter β by $\beta\,\beta_0$ or $\beta + \beta_0$ depending on the outlier type.

$$\begin{aligned} f(y_s|\,\theta) = \; & D(s)\,[(s-1)F^{s-2}(1-F)^{m-s}F^{\star}f \\ & +(m-s)F^{s-1}(1-F)^{m-s-1}(1-F^{\star})f \\ & +F^{s-1}(1-F)^{m-s}f^{\star}], \end{aligned} \tag{13}$$

where

$$D(s)\,=\,\binom{m-1}{s-1} \tag{14}$$

A. Outliers of type $\beta\beta_0$

The Y_s density function, in the presence of a single outlier of type $\beta\beta_0$, in the Gomp$(\alpha,\, \beta)$ case may be received through the substituting of (1) and (2) for f and F in (13). The f^* and F^* values presented by (1) and (2), after the replacement of β by $\beta\beta_0$. It is possible to simplify the density function implementing the pdf $g_1(y_2|\alpha, \beta)$, where the cdf $G_1(y_s|\alpha, \beta)$ is given as follows:

$$\begin{aligned} g_1(y_s|\, \alpha,\, \beta) \,=\, & D(s)\, \alpha\, \beta\, e^{\alpha y_s}\Big[(m+\beta_0-s)\sum_{j=0}^{s-1} A_{1j}(y_s) \\ & + (s-1)\sum_{j=0}^{s-2} A_{2j}(y_s)\Big], \quad y_s \,>\, 0, \end{aligned} \tag{15}$$

where

$$\begin{aligned} A_{1j}(y_s) =\; & a_{1j}(s)\exp\{-\beta\omega_j(s)\phi(y_s;\, \alpha)\}, \\ A_{2j}(y_s) =\; & a_{2j}(s)\Big[\exp\{-\beta\,\omega_{1j}(s)\phi(y_s;\, \alpha)\} \\ & -\exp\{-\beta\,\omega_{j+1}(s)\phi(y_s;\, \alpha)\}\Big], \\ \phi(y_s;\, \alpha) =\; & (e^{\alpha y_s}-1) \\ \omega_j(s) =\; & m\,-\,s\,+\,\beta_0\,+\,j, \\ \omega_{1j}(s) =\; & m\,-\,s\,+\,j\,+\,1 \end{aligned} \tag{16}$$

and for $\ell = 1,\, 2$,

$$a_{\ell\, j}(s)\,=\,(-1)^j\binom{s-\ell}{j}. \tag{17}$$

and the pdf $g_1(y_s|\, \alpha,\, \beta)$ the cdf $G_1(y_s|\, \alpha,\, \beta)$ is given by

$$\begin{aligned} G_1(y_s|\, \alpha,\, \beta) =\, & D(s)\,\Big[(m+\beta_0-s)\sum_{j=0}^{s-1} A^*_{1j}(y_s) \\ & + (s-1)\sum_{j=0}^{s-2} A^*_{2j}(y_s)\Big], \quad y_s \,>\, 0 \end{aligned} \tag{18}$$

where

$$\begin{aligned} A^*_{1j}(y_s) \;=\; & \frac{a_{1j}(s)}{\omega_j(s)}F(y_s;\, \alpha,\, \beta\omega_j(s)), \\ A^*_{2j}(y_s) \;=\; & \frac{a_{2j}(s)}{\omega_{1j}(s)}F(y_s;\, \alpha,\, \beta\omega_{1j}(s)) \\ & -\frac{a_{2j}(s)}{\omega_{j+1}(s)}F(y_s;\, \alpha,\, \beta\omega_{j+1}(s)). \end{aligned} \tag{19}$$

The Bayesian predictive density of y_s, $s = 1,\, 2,\, \cdots,\, m$ given $\underline{x}$ is represented by

$$g_1^*(y_s|\underline{x}) = \int_0^{\infty}\int_0^{\infty} g_1(y_s|\, \alpha,\, \beta)\, \pi^*(\alpha,\, \beta|\, \underline{x})\, d\alpha\, d\beta. \tag{20}$$

The Bayesian predictive distribution function of y_s, $s = 1,\, 2,\, \cdots,\, m$ given $\underline{x}$, α and β is given by

$$G_1^*(y_s\,|\underline{x}) = \int_0^{\infty}\int_0^{\infty} G_1(y_s|\, \alpha,\, \beta)\, \pi^*(\alpha,\, \beta|\, \underline{x})\, d\alpha\, d\beta. \tag{21}$$

Supposing that $\{(\alpha_i,\, \beta_i);\, i \,=\, 1,\, 2,\, \cdots, M\}$ are MCMC samples received from $\pi^*(\alpha,\, \beta|\, \underline{x})$, it is possible to get the simulation consistent estimators of $g_1^*(y_s|\underline{x})$ and $G^*(y_s|\underline{x})$ can be obtained as

$$\hat{g}_1^*(y_s\,|\underline{x}) \,=\, \sum_{i=1}^{M} g_1(y_s|\,\alpha_i,\,\beta_i)\,h_i \tag{22}$$

and

$$\hat{G}_1^*(y_s\,|\underline{x}) \,=\, \sum_{i=1}^{M} G_1(y_s|\,\alpha_i,\,\beta_i)\,h_i \tag{23}$$

where

$$h_i \,=\, \frac{h_3(\alpha_i,\,\beta_i)}{\sum\limits_{i=1}^{M} h_3(\alpha_i,\,\beta_i)}; \quad i = 1,\,2,\,\cdots,\,M. \tag{24}$$

A $(1-\tau)\,100\,\%$ Bayesian prediction interval for Y_s is as follows: $P[L(\underline{x}) \,\le\, Y_s \,\le\, U(\underline{x})] \,=\, 1-\tau$, where $L(\underline{x})$ and $U(\underline{x})$ are the lower and the upper bounds for y_s, $s = 1, 2, \cdots, m$. Thus, equating of (23) $1-\frac{\tau}{2}$ and $\frac{\tau}{2}$, enables to get the following:

$$P[Y_s \,\ge\, L(\underline{x})|\,\underline{x}] = 1 - \frac{\tau}{2} \,\Rightarrow\, \hat{G}_1^*(L(\underline{x})|\,\underline{x}) \,=\, \frac{\tau}{2} \tag{25}$$

and

$$P[Y_s \,\le\, U(\underline{x})|\,\underline{x}] = \frac{\tau}{2} \,\Rightarrow\, \hat{G}_1^*(U(\underline{x})|\,\underline{x}) \,=\, 1 - \frac{\tau}{2}. \tag{26}$$

B. Type $\beta + \beta_0$ Outliers

The y_s density function, in the presence of a single outlier of type $\beta + \beta_0$, in the Gomp$(\alpha,\,\beta)$ case, can be received through the substituting of (1) and (2) for F and f in (3). The F^* and f^* are presented by (1) and (2) after the replacement of β by $\beta + \beta_0$. Consequently, the density begins to form:

$$\begin{aligned} g_2(y_s|\,\alpha,\,\beta) = D(s)\,e^{\alpha\,y_s}\,\Big[&(\beta\,(m-s+1)+\beta_0)\sum_{j=0}^{s-1} B_{1j}(y_s) \\ &+ \beta(s-1)\sum_{j=0}^{s-2} B_{2j}(y_s)\Big],\; y_s > 0, \end{aligned} \tag{27}$$

where

$$\begin{aligned} B_{1j}(y_s) =\;& a_{1j}(s)\exp\left\{-\left[\beta\,\omega_{1j}(s)+\beta_0\right]\phi(y_s;\,\alpha)\right\} \\ B_{2j}(y_s) =\;& a_{2j}(s)\Big[\exp\left\{-\beta\,\omega_{1j}(s)\phi(y_s;\,\alpha)\right\} \\ &-\;\exp\left\{-\left[\beta\,\omega_{1(j+1)}(s)+\beta_0\right]\phi(y_s;\,\alpha)\right\}\Big], \end{aligned} \tag{28}$$

$\phi(y_s;\,\alpha)\,\omega_{1j}(s)$ are given in (16) and $a_{\ell j}(s)$, $a_2j(s)$ is given for $\ell = 1,\,2$, respectively, by (17).
The cdf corresponding to the pdf $g_2(y_s|\,\alpha,\,\beta)$ is presented by

$$\begin{aligned} G_2(y_s|\,\alpha,\,\beta) = D(s)\,\Big[&(\beta\,(m-s+1)+\beta_0)\sum_{j=0}^{s-1} B_{1j}^*(y_s) \\ &+ \beta(s-1)\sum_{j=0}^{s-2} B_{2j}^*(y_s)\Big],\; y_s > 0, \end{aligned} \tag{29}$$

where

$$\begin{aligned} B_{1j}^*(y_s) =\;& \frac{a_{1j}(s)}{\beta\,\omega_{1j}(s)+\beta_0}F(y_s;\,\alpha,\,\beta\,\omega_{1j}(s)+\beta_0), \\ B_{2j}^*(y_s) =\;& \frac{a_{2j}(s)}{\beta\,\omega_{1j}(s)}F(y_s;\,\alpha,\,\beta\,\omega_{1j}(s)) \\ -\;& \frac{a_{2j}(s)}{\beta\,\omega_{1(j+1)}(s)+\beta_0}F(y_s;\,\alpha,\,\beta\,\omega_{1(j+1)}(s)+\beta_0), \end{aligned} \tag{30}$$

where $F(y_s;\,\alpha,\,\beta\,m+\beta_0)$ is given by(2).

The Bayesian predictive distribution function of y_s, $s = 1,\,2,\,\cdots,\,m$ given $\underline{x}$, α and β is given by

$$g_2^*(y_s|\underline{x}) \,=\, \int_0^\infty \int_0^\infty g_2(y_s|\,\alpha,\,\beta)\,\pi^*(\alpha,\,\beta|\,\underline{x})\,d\alpha\,d\beta, \tag{31}$$

and the predictive cdf of y_s, $G_2^*(y_s|\underline{x})$ is given by

$$G_2^*(y_s|\underline{x}) \,=\, \int_0^\infty \int_0^\infty G_2(y_s|\,\alpha,\,\beta)\,\pi^*(\alpha,\,\beta|\,\underline{x})\,d\alpha\,d\beta, \tag{32}$$

where $G_2(y_s|\,\alpha,\,\beta)$ is given by (29) and $\pi^*(\alpha,\,\beta|\,\underline{x})$ is given by (9). It is evident that it is impossible to express (31) and (32) in closed form. Therefore, they cannot be analytically evaluated.
The use of MCMC samples $\{(\alpha_i,\,\beta_i),\,i \,=\, 1,\,2,\,\cdots,\,M\}$, enable the obtaining of $g_2^*(y_s|\underline{x})$ and $G_2^*(y_s|\underline{x})$ simulation consistent estimator, as follows:

$$\hat{g}_2^*(y_s|\underline{x}) \,=\, \sum_{i=1}^{M} g_2(y_s|\,\alpha_i,\,\beta_i)\,h_i, \tag{33}$$

and

$$\hat{G}_2^*(y_s|\underline{x}) \,=\, \sum_{i=1}^{M} G_2(y_s|\,\alpha_i,\,\beta_i)\,h_i, \tag{34}$$

Where h_i is given by (24). It is essential to highlight that it is possible to use the same MCMC samples $\{(\alpha_i,\,\beta_i),\,i \,=\, 1,\,2,\,\cdots,\,M\}$, to compute $\hat{g}_2^*(y_s|\underline{x})$ and $\hat{G}_2^*(y_s|\underline{x})$ for all y_s. Also, A $(1-\tau)100\%$ Bayesian prediction intervals for is $P[L(\underline{x}) \,\le\, Y_s \,\le\, U(\underline{x})] \,=\, 1-\tau$ where $L(\underline{x})$ and $U(\underline{x})$ are lower and upper y_s Bayesian prediction bounds. Hence, it is possible to get the lower and upper Bayesian prediction bounds, $L(\underline{x})$ and $U(\underline{x})$, for $y_s, s = 1, 2, \cdot, m$ through solving the following two nonlinear equations.

$$P[Y_s \,\ge\, L(\underline{x})|\,\underline{x}] = 1 - \frac{\tau}{2} \,\Rightarrow\, \hat{G}_2^*(L(\underline{x})|\,\underline{x}) \,=\, \frac{\tau}{2} \tag{35}$$

and

$$P[Y_s \,\le\, U(\underline{x})|\,\underline{x}] = \frac{\tau}{2} \,\Rightarrow\, \hat{G}_2^*(U(\underline{x})|\,\underline{x}) \,=\, 1 - \frac{\tau}{2}. \tag{36}$$

It is possible to solve the two nonlinear equations (35) and (36) through the use of an iterative method to receive the lower and upper Bayesian prediction bounds for $y_s; s = 1, 2, \cdots, m$.

V. Numerical Example

Example 1. This example shows a doubly Type-II censored sample, $x_{(s+1)}, x_{(s+2)}, \cdots, x_{(r)}$, that is received through the application of the following steps:

1 – For the hyperparameters given values $\eta_1 = 1.2$ and $\gamma_1 = 1.8$ a generated value of $\alpha = 0.860986$ is received from the prior distribution with pdf (6).

2 – For the hyperparameters given values $\eta_2 = 1.4$ and $\gamma_2 = 1.7$ a generated value of $\beta = 0.409442$ is received from the prior distribution with pdf (7).

3 – The use of the generated values of α and β from two prior steps, enables to generate a sample of size $n = 30$ from the Gomp(α, β) distribution with pdf, that is represented by (2).

4 – The application of some sorting routine, assists in obtaining a doubly Type-II censored different value sample of size $r = 20, 25, 30$ and $k = 0, 5, 10$ from the Gomp(α, β) distribution, where the deferment value of r and k is presented in Tables I, II and III.

5 – Generate (α_i, β_i), $i = 1, 2, \cdots, M$, through the use of MCMC shown in Algorithm 1.

6 – The above generated doubly Type-II censored size $(r-s)$ sample, the 95 % Bayesian prediction links to the future ordered values, $y_{(1)}, y_{(2)}, \cdots, y_{(m)}$, $m = 5$ in the single types $\beta\,\beta 0$ outliers, enable a numerical calculation through solving the equations (25) and (26).

Let us assume that we have one more size $m = 5$ sample in the presence of a single outlier of type $\beta\,\beta_0$. Hence, for the given β_0 values we seek to receive 95% Bayesian prediction bounds for y_1 to y_5 of the failure future sample times. Tables I, II and III represents these bounds with the corresponding β_0 values.

Example 2. The 95% Bayesian prediction interval for a future unobserved y_1 to y_5, which are the failure times in the future size 5 sample in the presence of a single outlier of type $\beta + \beta_0$ can be obtained on the basis of a generated doubly Type-II censored sample of size$n$ from the Gomp (α, β) distribution. Same different η_1, γ_1, η_2, γ_1 hyper-parameter values and the same data set is presented in Example 1. Hence, these bounds with the corresponding $n = 30$, $r = 20, 25, 20$ and $k = n - r$ and β_0 values are shown in Tables IV, V and VI.

VI. Conclusion

The study investigated and discussed the single $\beta\,\beta_0$ and $\beta + \beta_0$ type outliers through the application of the predictive distribution function. Hence, the Bayesian prediction intervals in the case of future homogeneous case observations can be received by $\beta_0 = 1$ in (18) or $\beta_0 = 0$ in (29).
However, it is impossible in the no outlier case. The Gibbs sampling technique was applied to generate MCMC samples. Afterwards, the importance sampling methodology was used to compute the Bayesian prediction problems in the presence of a single outlier of both type. It is essential to highlight that the Bayesian prediction intervals are shorter for y_1 and larger for the Bayesian prediction intervals for y_5 due to the increase of β_0 value.

TABLE I. 95 % Bayesian prediction intervals for $y_1, \cdots, y_5$ in the presence of a single outlier of type $\beta\,\beta_0$, where $n = 30$, $r = 20$, $k = 10$. Note: Obs. is observations PP is point predictors, LB is Lower Bound, UB is Upper Bound, CP is Coverage Percentages.

β_0	Obs	y_1	y_2	y_3	y_4	y_5
1	PP	0.492721	0.933387	1.35569	1.80134	2.37285
	LB	0.017663	0.177097	0.460092	0.824819	1.29551
	UB	1.39901	1.88431	2.30933	2.76994	3.43655
	Length	1.38135	1.70721	1.84924	1.94513	2.14104
	CP	95.77 %	95.62 %	95.03 %	94.79 %	93.86 %
2	PP	0.428709	0.836901	1.24704	1.69552	2.285
	LB	0.014736	0.151347	0.403924	0.743443	1.19871
	UB	1.2507	1.73927	2.1873	2.68116	3.38701
	Length	1.23597	1.58792	1.78338	1.93772	2.1883
	CP	95.17 %	95.84 %	95.91 %	96.12 %	95.65 %
3	PP	0.379907	0.77606	1.19357	1.65873	2.26774
	LB	0.012642	0.13457	0.370458	0.700365	1.15708
	UB	1.13273	1.66262	2.15497	2.67353	3.38666
	Length	1.12009	1.52805	1.78451	1.97316	2.22958
	CP	94.19 %	95.73 %	96.4%	96.77 %	96.31 %
4	PP	0.34137	0.735004	1.16389	1.64286	2.26251
	LB	0.011069	0.122535	0.34753	0.6731	1.13514
	UB	1.03625	1.62423	2.14778	2.67299	3.38665
	Length	1.02518	1.5017	1.80025	1.99989	2.25152
	CP	92.88 %	95.61 %	96.64 %	97.25 %	96.6 %
5	PP	0.310111	0.705889	1.14598	1.63502	2.2605
	LB	0.009844	0.113362	0.33055	0.654265	1.12263
	UB	0.955644	1.60625	2.14633	2.67296	3.38665
	Length	0.945801	1.49289	1.81578	2.01869	2.26402
	CP	91.31 %	95.49 %	96.86 %	97.55 %	96.77 %

TABLE II. 95 % Bayesian prediction intervals for $y_1, \cdots, y_5$ in the presence of a single outlier of type $\beta\,\beta_0$, where $n = 30$, $r = 25$, $k = 5$. Note:Obs. is observations PP is point predictors, LB is Lower Bound, UB is Upper Bound, CP is Coverage Percentages.

β_0	Obs	y_1	y_2	y_3	y_4	y_5
1	PP	0.480366	0.912016	1.3272	1.76659	2.33151
	LB	0.017118	0.171888	0.447624	0.804559	1.26715
	UB	1.3691	1.84794	2.26819	2.72426	3.38507
	Length	1.35198	1.67605	1.82057	1.9197	2.11793
	CP	95.69 %	95.69 %	95.12 %	95.04 %	94.49 %
2	PP	0.417723	0.817269	1.22021	1.66216	2.24463
	LB	0.014281	0.146862	0.392804	0.724791	1.17186
	UB	1.22303	1.7047	2.14747	2.63631	3.33594
	Length	1.20875	1.55784	1.75466	1.91152	2.16407
	CP	94.88 %	95.76 %	95.86 %	96.25 %	96.03 %
3	PP	0.370002	0.757583	1.16762	1.6259	2.22759
	LB	0.012251	0.130561	0.360161	0.682593	1.13091
	UB	1.10696	1.62904	2.11549	2.62875	3.33559
	Length	1.09471	1.49848	1.75533	1.94616	2.20467
	CP	94. %	95.41 %	96.21 %	96.84 %	96.5 %
4	PP	0.332341	0.717339	1.13846	1.61028	2.22243
	LB	0.010726	0.118871	0.337806	0.655895	1.10933
	UB	1.01211	1.59117	2.10838	2.62822	3.33558
	Length	1.00139	1.4723	1.77057	1.97233	2.22625
	CP	92.5 %	95.15 %	96.33 %	97.33 %	96.79 %
5	PP	0.30181	0.688818	1.12087	1.60256	2.22046
	LB	0.009539	0.109964	0.321256	0.637456	1.09703
	UB	0.93294	1.57343	2.10694	2.62818	3.33558
	Length	0.923401	1.46347	1.78568	1.99073	2.23855
	CP	90.53 %	95.11 %	96.51 %	97.58 %	96.94 %

Acknowledgment

The authors extend their appreciation to the Deanship of Scientific research at Majmaah University for funding this work through research project No. R-1441-85.

References

[1] JC. Ahuja and SW Nash, *The Generalized Gompertz-Verhulst Family of Distributions*, Sankhya, 1967.

[2] J. Ali, T. Saeid, and A. Morad, *The beta-Gompertz distribution*, Revista Colombiana de Estadistica, 2014.

TABLE III. 95 % BAYESIAN PREDICTION INTERVALS FOR $y_1, \cdots, y_5$ IN THE PRESENCE OF A SINGLE OUTLIER OF TYPE $\beta\,\beta_0$, WHERE $n = 30$, $r = 30$, $k = 0$. NOTE:OBS. IS OBSERVATIONS PP IS POINT PREDICTORS, LB IS LOWER BOUND, UB IS UPPER BOUND, CP IS COVERAGE PERCENTAGES.

β_0	Obs	y_1	y_2	y_3	y_4	y_5
1	PP	0.475422	0.902867	1.31418	1.74964	2.30966
	LB	0.016929	0.170029	0.442908	0.796326	1.25459
	UB	1.35561	1.8302	2.24682	2.69903	3.35434
	Length	1.33868	1.66017	1.80392	1.9027	2.09975
	CP	95.56 %	95.69 %	95.21 %	95.11 %	94.63 %
2	PP	0.413396	0.809015	1.20817	1.64613	2.22353
	LB	0.014124	0.145269	0.388644	0.717329	1.16018
	UB	1.21088	1.68822	2.12713	2.61182	3.30561
	Length	1.19675	1.54295	1.73849	1.89449	2.14543
	CP	94.81 %	95.63 %	95.72 %	96.19 %	95.98 %
3	PP	0.366148	0.7499	1.15607	1.61019	2.20664
	LB	0.012116	0.129143	0.356335	0.675541	1.11961
	UB	1.09588	1.61323	2.09543	2.60433	3.30526
	Length	1.08376	1.48409	1.73909	1.92878	2.18565
	CP	93.81 %	95.25 %	95.95 %	96.84 %	96.49 %
4	PP	0.328865	0.710044	1.12718	1.59472	2.20152
	LB	0.010608	0.117579	0.334211	0.649105	1.09822
	UB	1.00191	1.57569	2.08838	2.6038	3.30526
	Length	0.991305	1.45811	1.75417	1.95469	2.20703
	CP	92.35 %	94.94 %	96.15 %	97.25 %	96.81 %
54	PP	0.298642	0.681802	1.10976	1.58707	2.19956
	LB	0.009434	0.108767	0.317832	0.630847	1.08604
	UB	0.923483	1.55811	2.08696	2.60376	3.30526
	Length	0.914049	1.44934	1.76912	1.97292	2.21922
	CP	90.22 %	94.88 %	96.31 %	97.43 %	96.93 %

TABLE IV. 95 % BAYESIAN PREDICTION INTERVALS FOR $y_1, \cdots, y_5$ IN THE PRESENCE OF A SINGLE OUTLIER OF TYPE $\beta + \beta_0$, WHERE $n = 30$, $r = 20$, $k = 10$. NOTE: OBS. IS OBSERVATIONS PP IS POINT PREDICTORS, LB IS LOWER BOUND, UB IS UPPER BOUND, CP IS COVERAGE PERCENTAGES.

β_0	Obs	y_1	y_2	y_3	y_4	y_5
0	PP	0.314436	0.624792	0.947281	1.3117	1.80927
	LB	0.010012	0.103158	0.280504	0.530439	0.8847
	UB	0.966839	1.36921	1.7411	2.15974	2.78612
	Length	0.956827	1.26605	1.46059	1.6293	1.90142
	CP	95.48 %	95.68 %	95.32 %	95.39 %	95.04 %
1	PP	0.249322	0.524206	0.833376	1.20209	1.72088
	LB	0.00759	0.081351	0.230135	0.452727	0.787286
	UB	0.792205	1.20258	1.61103	2.07199	2.7386
	Length	0.784615	1.12123	1.3809	1.61926	1.95131
	CP	94.01 %	94.96 %	95.6 %	96.61 %	96.6 %
2	PP	0.206843	0.474838	0.794469	1.17926	1.71243
	LB	0.006111	0.069507	0.205725	0.420648	0.757922
	UB	0.67229	1.14618	1.59745	2.07048	2.73858
	Length	0.666179	1.07668	1.39172	1.64983	1.98066
	CP	91.36%	94.35%	95.9 %	97.18 %	97.02 %
3	PP	0.176862	0.446821	0.777449	1.17201	1.71067
	LB	0.005115	0.06181	0.190618	0.402904	0.746022
	UB	0.584482	1.12971	1.59639	2.07045	2.73858
	Length	0.579367	1.0679	1.40577	1.66755	1.99256
	CP	87.7%	94.34%	96.26%	97.31%	97.2%
4	PP	0.154538	0.429346	0.768791	1.16911	1.71014
	LB	0.004398	0.056299	0.180133	0.391883	0.740829
	UB	0.517254	1.12552	1.59631	2.07045	2.73858
	Length	0.512856	1.06922	1.41618	1.67857	1.99775
	CP	84.05%	94.45%	96.45%	97.45%	97.28 %

TABLE V. 95 % BAYESIAN PREDICTION INTERVALS FOR $y_1, \cdots, y_5$ IN THE PRESENCE OF A SINGLE OUTLIER OF TYPE $\beta + \beta_0$, WHERE $n = 30$, $r = 25$, $k = 5$. NOTE: OBS. IS OBSERVATIONS PP IS POINT PREDICTORS, LB IS LOWER BOUND, UB IS UPPER BOUND, CP IS COVERAGE PERCENTAGES.

β_0	Obs	y_1	y_2	y_3	y_4	y_5
0	PP	0.310392	0.617041	0.935936	1.29655	1.78929
	LB	0.009873	0.101746	0.276772	0.523637	0.873873
	UB	0.955124	1.35334	1.72163	2.1364	2.75727
	Length	0.945251	1.25159	1.44485	1.61277	1.88339
	CP	95.48%	95.69%	95.34%	95.48%	95.16%
1	PP	0.246376	0.518007	0.823625	1.18832	1.70186
	LB	0.007495	0.080324	0.227241	0.447124	0.777809
	UB	0.783238	1.18895	1.59301	2.04949	2.71016
	Length	0.775743	1.10863	1.36578	1.60236	1.93235
	CP	93.86%	94.87%	95.47%	96.52%	96.53%
2	PP	0.204538	0.469253	0.785086	1.16562	1.69343
	LB	0.00604	0.06866	0.203175	0.41544	0.748708
	UB	0.665036	1.13293	1.57938	2.04795	2.71014
	Length	0.658996	1.06427	1.37621	1.63251	1.96143
	CP	91.04 %	94.21%	95.75%	97.01%	96.97 %
3	PP	0.174974	0.441532	0.768182	1.15839	1.69166
	LB	0.005059	0.061072	0.188266	0.397885	0.736867
	UB	0.578397	1.11642	1.5783	2.04793	2.71014
	Length	0.573339	1.05535	1.39004	1.65004	1.97327
	CP	87.33%	94.02%	96.06%	97.16%	97.14%
4	PP	0.152942	0.424219	0.759568	1.15549	1.69113
	LB	0.004351	0.055635	0.17791	0.386966	0.731677
	UB	0.512017	1.11217	1.57822	2.04793	2.71014
	Length	0.507666	1.05654	1.40031	1.66096	1.97846
	CP	3.69%	94.13%	96.27%	97.27%	97.24 %

TABLE VI. 95 % BAYESIAN PREDICTION INTERVALS FOR $y_1, \cdots, y_5$ IN THE PRESENCE OF A SINGLE OUTLIER OF TYPE $\beta + \beta_0$, WHERE $n = 30$, $r = 30$, $k = 0$. NOTE: OBS. IS OBSERVATIONS PP IS POINT PREDICTORS, LB IS LOWER BOUND, UB IS UPPER BOUND, CP IS COVERAGE PERCENTAGES.

β_0	Obs	y_1	y_2	y_3	y_4	y_5
0	PP	0.309735	0.615638	0.933667	1.29322	1.78437
	LB	0.009855	0.10156	0.276229	0.522522	0.871834
	UB	0.952855	1.34988	1.71698	2.13035	2.74903
	Length	0.943	1.24831	1.44075	1.60783	1.8772
	CP	95.48 %	95.67 %	95.32 %	95.46 %	95.14 %
1	PP	0.245765	0.516724	0.821549	1.18522	1.69719
	LB	0.007478	0.080147	0.226737	0.446102	0.77594
	UB	0.781161	1.18581	1.58871	2.04372	2.70209
	Length	0.773683	1.10566	1.36198	1.59762	1.92615
	CP	93.81%	94.77%	95.43%	96.46%	96.53%
2	PP	0.203983	0.46808	0.783138	1.16263	1.68881
	LB	0.006025	0.068498	0.202712	0.414491	0.74694
	UB	0.66315	1.13002	1.57519	2.0422	2.70207
	Length	0.657125	1.06152	1.37248	1.62771	1.95513
	CP	90.95%	94.14%	95.69%	96.96%	96.95%
3	PP	0.174471	0.44044	0.766305	1.15544	1.68705
	LB	0.005045	0.060923	0.187833	0.396987	0.735156
	UB	0.576681	1.11363	1.57412	2.04218	2.70207
	Length	0.571636	1.05271	1.38629	1.6452	1.96692
	CP	87.27%	93.94%	96.01%	97.09%	97.13%
4	PP	0.152484	0.423186	0.757732	1.15256	1.68652
	LB	0.004339	0.055497	0.177501	0.386105	0.729999
	UB	0.510446	1.10943	1.57404	2.04218	2.70207
	Length	0.506108	1.05393	1.39654	1.65608	1.97207
	CP	83.63%	94.05%	96.22%	97.2%	97.23%

[3] E.K. AL-Hussaini, G.R. AL-Dayian, and S.A. Adham, *On Finite Mixture of Two-Component Gompertz Lifetime Mode*, J. Statist. Comput. Simul., 2000.

[4] N. Balakrishnan, and R.S. Ambagaspitiya *Relationships among moments of order statistics in samples from two related outlier models and some applications*, Comm. Statist. Theory Methods, 1988.

[5] V. Barnett, and T. Lewis, *Outliers in Statistical Data*, Wiley, New York, 1984.

[6] L. Gavrilov, and N. Gavrilova, *The biology of Life Span: A Quantitative Approach*, Chur: Harwood 1991.

[7] M. Garg, B. Rao, and C. Redmond, *Maximum-likelihood estimation of the parameters of the Gompertz survival function. Journal of the Royal Statistical Society*, Journal of the Royal Statistical Society. Series C (Applied Statistics), 1970.

[8] A. El-Gohary, A. Alshamrani, and A. Al-Otaibi, *The generalized Gompertz distribution*, Applied Mathematical Modeling, 2013.

[9] Z.F. Jaheen, *Bayesian prediction under a mixture of two-component Gompertz lifetime model*, Test, 2003a.

[10] Z.F. Jaheen, *A Bayesian analysis of record statistics from the Gompertz model*, Applied Mathematics and Computation, 2003b.

[11] A. Marshall, and I. Olkin, *Life Distributions.*, Springer, 2007.

[12] N. Metropolis, A. W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, *Equations of state calculations by fast computing machine*, The Journal of Chemical Physics, 1953.

[13] A.F. Niazi Ali, *Prediction intervals based on Gompertz doubly censored*

data, Comput Stat, 2016.

[14] C.P. Robert, and G. Casella, *Moizte Ccir-io Stciristicai Methocis*, New York: Springer-Verlag, 1999.

[15] A.M. Sarhan, and J. Apaloo, *Exponentiated modified Weibull extension distribution*, Reliability Engineering and System Safety, 2013.

[16] J. Vaupel, *How change in age-specific mortality affects life expectancy*, Population Studies, 1986

[17] W. Wetterstrand, *Parametric models for life insurance mortality data: Gompertz's law over time*, Transactions of the Society of Actuaries, 1981.

[18] F. Willekens, *Gompertz in context: the Gompertz and related distributions. In Forecasting Mortality in Developed Countries - Insights from a Statistical, Demographic and Epidemiological Perspective*, European Studies of Population, vol. 9, edited by E. Tabeau, A. van den Berg Jeths and C. Heathcote, Springer, 2002.

[19] W. Willemse, and H. Koppelaar, *Knowledge Elicitation of Gompertz' Law of Morality*, Scandinavian Actuarial Journal, 2000.

[20] M.M. El-Din, M. Nagy, and M.H. Abu-Moussa, *Estimation and Prediction for Gompertz Distribution Under the Generalized Progressive Hybrid Censored Data*, Annals of Data Science, 2019.

[21] M.M. Mohie El-Din, Y. Abdel-Aty, and M.H. Abu-Moussa, *Statistical inference for the Gompertz distribution based on Type-II progressively hybrid censored data*, Commun Stat Simul Comput, 2017.

[22] M.M. Mohie El-Din and M.H. Abu-Moussa, *statistical inference and prediction for the Gompertz distribution based on multiply Type-I censored data*, J Egypt Math Soc, 2018.

Code Readability Management of High-level Programming Languages: A Comparative Study

Muhammad Usman Tariq[1]
Abu Dhabi School of Management
Abu Dhabi, UAE

Muhammad Bilal Bashir[2], Muhammad Babar*[3], Adnan Sohail[4]
Computing & Technology Department
IQRA University, Islamabad, Pakistan

***Abstract*—Quality can never be an accident and therefore, software engineers are paying immense attention to produce quality software product. Source code readability is one of those important factors that play a vital role in producing quality software. The code readability is an internal quality attribute that directly affects the future maintenance of the software and re-usability of same code in similar other projects. Literature shows that readability does not just rely on programmer's ability to write tidy code but it also depends on programming language's syntax. Syntax is the most visible part of any programming language that directly influence the readability of its code. If readability is a major factor for a given project, the programmers should know about the language that they shall choose to achieve the required level of quality. For this we compare the readability of three most popular high-level programming languages; Java, C#, and C++. We propose a comprehensive framework for readability comparison among these languages. The comparison has been performed on the basis of certain readability parameters that are referenced in the literature. We have also implemented an analysis tool and performed extensive experiments that produced interesting results. Furthermore, to judge the effectiveness of these results, we have performed statistical analysis using SPSS (Statistical Package for Social Sciences) tool. We have chosen the Spearman's correlation ad Mann Whitney's T-test for the same. The results show that among all three languages, Java has the most readable code. Programmers should use Java in the projects that have code readability as a significant quality requirement.**

***Keywords*—Source code; high-level programming languages; Java; C++; C#; code readability; code readability index**

I. Introduction

Software engineering is different in nature as compared to other engineering domains. Products may remain in use even if there are some imperfections in them. But a software product may go through several revisions even after development is completed until software becomes faults free. Otherwise customer may not accept and use it. Customers these days are very smart and want to know what is going inside the software and what does affect the future maintenance and cost.

Software go through several updates after the first version due to some reasons; a feature was not implemented that was required, a feature was incorrectly implemented, or a new feature is now required. This is known as maintenance and research shows that around 70% of the product cost is spent on the maintenance [2] as shown in Fig. 1. Software engineers need to ensure that the software they produce is easy to maintain. There are many factors that affect software maintainability and source code readability is one of them. Readability is how quickly a reader can read and understand the written text. Elements that make the text difficult to read and understand include; long lines, insufficient contrast, and long paragraph with no segmentation.

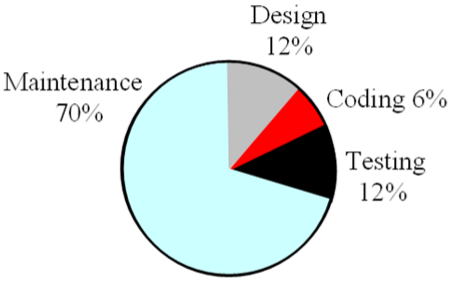


Fig. 1. Cost Distribution among Software Process Activities [12]

In a software product, readability means the ability to read documentation and source code [10]. The documentation serves as the means of communication among the stakeholders. But the research shows that agile teams focus on working software as compared to the documentation while communicating with the clients [10]. Collection of computer instructions that are written in high-level programming language is called source code. Source code is the significant part of software readability in terms of re-usability, cost, maintenance, and robustness. Software industry is facing problems to minimize the software development cost, which is affected by many factors. Researchers are trying to identify those factors and ways to eliminate or at least reduce their impact to reduce the overall cost. According to Collar et al. [11] improved readability saves developer's time while reading the code that eventually helps in bringing down the overall development cost. Readability is important not only during development time to improve software quality [1] but also during maintenance because reading the code is the first stage of maintenance [3]. Research also shows that the maintainability of a software is measured by the readability and understand-ability of code. [12].

If for a given project, project manager foresees that a large number of programmers will be required, programmers are geographically distributed, programmers will be changing over the period of time, new programmers will be hired, or customers will change the requirements then code readability becomes a major concern. Generally code readability is calculated using proportion between number of lines and the comments that are written for the programmer. The project manager should select a programming language, which is not only suitable for project's functional requirements but should also offer required level of readability. This selection is vital because the correct selection will positively affect the quality of the software.

In this research we have conducted a comparative study on readability of high-level programming languages. We have chosen Java, C++, and C# for this purpose. According to the TIOBE programming community index [15], Java, C++ and C# are among the top five high-level programming languages. These languages are maximally used, so we have computed the readability value of these languages. For this first we have devised a comprehensive framework and used it for the analysis. The analysis is three-fold, we have not only used general text readability indexes, code readability indexes, but also have included the expert opinions. The end results clearly shows that Java has been the best as far as readability is concerned among all.

Rest of the paper is organized as follows. Section II presents brief description on literature review of existing text readability assessment techniques. Section III covers all the proposed techniques for code readability analysis. In Section IV, we present our novel framework to perform comparative analysis among programming languages. Section V presents experiment details and results. We analyze results using statistical techniques in Section VI. Finally we conclude the discussion in Section VII and future directions in Section VIII.

II. Literature Review

In this section, we present literature review of readability metrics to assess the natural languages. Readability tests not only determine readability but also predict the reading ease. Most of the tests are language neutral but some of them are used for certain languages. We have used four natural language metrics for code readability assessment on the basis of their popularity and they are described in this section along with some others.

A. Coleman–Liau Index

Colman–Liau is a readability index similar to automated readability index (ARI) [16] but different from other indexes used to estimate the readability of text. This index is developed by Pahal et al. [3]. This index considers letters per word rather than text as a whole. It was used to calculate readability mechanically from samples of hard copy text. It does not require characters from words and it only calculates the length in characters. The formula of Coleman-Liau index is given below:

$$CLI = 0.0588L - 0.296S - 15.8$$

In the above mentioned equation "L" is average number of letters, whereas, "S" is average number of sentences.

B. SMOG

SMOG stands for "Simple Measure of Gobbledygook". McLaughlin [14] created this index in 1969 in article, SMOG Grading. It estimates the time (years) to read the text required by any person. As compared to other readability metrics, SMOG is better and provide more accurate results. SMOG metric is calculated with the following formula:

$$SMOG = 3 + SquarerootofPolysyllableCount$$

C. Flesch-Kincaid Readability Index

The Flesch-Kincaid [17] index is improved version of Flesch Reading Ease Readability Formula [3]. It checks the reading ease of the give text. If the value is high, it means the text readability is high. But if the value is low then it means text is difficult to read. The grade level is calculated with the following formula:

$$FKRI = 206.835 - 1.015\left(\tfrac{Totalwords}{TotalSentences}\right) - 84.6\left(\tfrac{TotalSyllables}{TotalWords}\right)$$

Shorter sentences and words give best results. The score between 60 and 69 is considered average readability while score between 0 and 29 is considered confusing for the reader. The complete list of values and their interpretations is provided in Table I.

TABLE I. Value Ranges and Description [17]

Score	Grade Level
90-100	Very Easy
80-89	Easy
70-79	Fairly Easy
60-69	Standard
50-59	Fairly Difficult
30-49	Difficult
0-29	Very Confusing

D. The Gunning's Fog Index

Gunning [18] propose this index and it is also known as FOG index in short. It can be calculated by using the following formula:

$$FOG = 0.4(ASL + PHW)$$

The average sentence length is added to the percentage of hard word (PHW). And average sentence length (ASL) is calculated by ratio of words count to the total number of sentences. Ideal score for FOG readability is 7 or 8 and if score goes higher than 12, it is considered as hard to read text.

E. The Automated Readability Index (ARI)

Senter [19] design automated readability index (ARI) test to access the understandability of text. Word difficulty and sentences are used in ARI. ARI calculate the readability value and output will be compared with grade level. Here is the formula of ARI:

$$ARI = 4.71\left(\frac{Characters}{Words}\right) + 0.5\left(\frac{Words}{Sentences}\right) - 21.43$$

Characters are the number of letters and numbers. Words are the number of words and spaces and sentences are the number of sentences.

III. Code Readability Indexes

The most important parameter of maintainable software is readability, because changes in the system are made through source code [3]. Less readable source code is harder to maintain than a code that is readable. Most of the time managers reject the code due to lack of code readability. In this section we present some code readability index that we find in the literature.

A. Deepa and Dua (2015)

Deepa and Dua [4] explain that readability depends upon simple sequences and unnecessary loops complicate the program. In this paper code readability is calculated on the basis of software developer judgment. Authors use two copies of the same program for their study. First copy of the program is less readable as proper indentation was not applied whereas the second copy was well formatted using a beautifier tool. Authors also propose a new metric for readability assessment. They perform experiments using novel readability metric and find out that the program written and formatted properly with the help of beautifier has more readability as compared to the other one. The metric that authors use have some parameters including; lines of code, line length, and number of comment lines, number of blank lines, number of lines after semicolon, number of spaces after directive statement and number of method.

B. Tashtoush (2013)

Tashtoush [5] develops an approach called “impact of programming features on code readability” (IPFCR). In this approach author studies the impact of various features and their effect on code readability. For evaluation he uses feature code readability tool (CRT). Author conducts the survey on a random number of expert programmers to access the level of impact. 25 readability features are proposed for survey; meaningful name, comments, spacing, indents, short scope, line length distribution, identifier name length, arithmetic formula, identifier frequency, if-else, nested if, switch, for loop, do while loop and nested loop [5]. Programmers evaluated features into positive and negative factors based on their understandability. The results are evaluated using SPSS statistical tool. ANOVA test is used to remove the biased from data. The top three features that come from survey were meaningful names, consistency and comments. And the lowest impact features were nested loops, arithmetic formula and recursive function. Some of them have neutral impact on readability.

C. Sivaprakasam and Sangeetha (2012)

Sivaprakasam and Sangeetha [7] have conducted a study that shows that readability has a global effect on software budget. In this paper authors define the relationship between software quality and source code readability. Mostly software metrics are used to measure the complexity of software. Authors have developed an automated readability tool, which is 80% more effective than human judgment. Authors have performed extensive experiments to evaluate the readability of code and for this they selected code snippets from the developed projects. The size of snippets is important because too small snippets may reflect incorrect or misleading scores. The scores authors have used range from 1 to 5 where 5 means more readable and 1 means least readable. Authors have ensured that all the snippets have some features including line length, number of character, identifier length indentation, loops and many other features. For a large number of experiments this technique is useful for conducting readability index.

D. Relf (2004)

Relf [8] examines in this paper that identifier naming standards that improve the code readability are acceptable by software professionals. Author claims that naming standards affect source code readability and that greatly impact code maintainability. To examine the impact of naming standards author collects 21 naming standards from research. These include multiple underscore characters, outside underscore character, numeric digits, naming convention anomaly, identifier encoding, short identifier name, long identifier name, number of words, class qualification, abstract words, constant qualification, numeric identifier name and some others. Author analyzes some codes written in ADA and Java programming languages and rates these programs on the basis of naming standards used from 1 to 5 (1 is strong acceptance and 5 is strong rejection). This study also states that expert programmers accept the naming standards more than the beginners.

E. DeYoung, Kampen, Topolski (1992)

An automated readability measure will be useful for developers during coding as it will continuously assessing their code and assisting them to improve. DeYoung et al. [9] examine the machine computable and human-judged program features. They identify that length of identifiers and are very useful in predicting code readability. Using analyzer generated quality of comments, logicality of control flow and meaningfulness of identifier names are studied to find out whether these predictors are worthy for readability estimation [9]. The proposed predictors increase the proportion of readability of judgments from 41% to 72%. Authors also claim that when logicality of control flow is added as a predictor, it produces better results as compared to human judgment but somehow these predictors are expensive to obtain.

F. Buse and Westley (2008)

Buse and Westley [2] perform a detailed empirical study to calculate readability of code. For this they have chosen 100 snippets and around 120 annotators that grade these snippets. The biggest issue in this research is that authors have used 19 parameters including line length, identifiers, identifier length, indentation, keywords, numbers, comments, periods, commas, spaces, parenthesis, arithmetic operators, comparison operators, assignment, branches, loops, blank lines, occurrences of any character and occurrences of any single identifier, which are difficult to calculate. From these parameters authors have constructed automated readability measurement and proved that it will be 80% more effective than human judgment. Furthermore, he discusses that how readability has potential for improving programming language design with respect to software quality. Authors also suggest to decrease the parameters for readability analysis and sets this as future work for their research.

G. Relf (2005)

Relf [6] describes a practical study to show whether coder increases the readability of his programs if he gets support from source code editor that provides vibrant responses on his identifier naming practices. Software coder should adopt a standard for software interface to gain benefits. This paper is useful for both student and professional software coder for maintaining the code and significant for the improvement of code readability. Author uses only one parameter for code readability that is identifier naming practices.

H. Daryl, Hindle, and Devanbu (2011)

Daryl et al. [13] propose to use entropy for predictive modeling approach. Authors study that whether size of the code impact the readability of the code or not. They have used six parameters including mathematical equations, average number of comments, and maximum indention, maximum word, maximum line length and maximum occurrence character in the code snippets. Author also used Halstead's metrics to find the size of code on the mean readability. For mean readability total number of operators and operands are combined and formulate the Halstead's metrics. For measuring the Entropy total number of tokens and unique token is counted. Also Entropy model improves the performance in term of prediction and readability but byte entropy does not improve the prediction.

IV. Framework for Comparative Analysis

In this section we present the framework we have proposed for performing the comparative analysis among the selected three programming languages (Java, C#, and C++).

The main objective of our work is to compare the readability of three of the top five most popular programming languages. In proposed framework we compare the human judgment with readability index: ARI (Automated Readability Index), SMOG, FOG, and FKG. The framework is presented in Fig. 2.

To perform comparison first we have to find the programing parameters that can affect the readability of source code. For this we select the constructs from the research work of Buse and Westley [2]. Second step is to compute the effect of these constructs on the readability of Java, C# and C++ languages. To calculate the effect, we have selected code snippets of Java, C# and C++ languages. After snippets selection, online survey is conducted, in which expert opinion is obtained and results are obtained for every selected programming construct. Selected snippets are measured with different readability indexes.

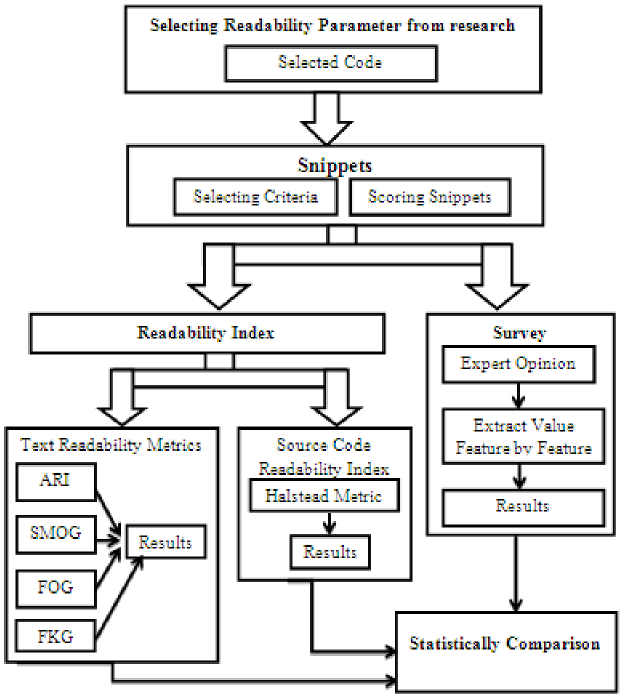


Fig. 2. The Proposed Framework

Text readability indexes include ARI (Automated Readability index), SMOG Fog, and Flesch Kincaid Grade level, while the code readability includes Halstead's complexity. The effect of readability by each construct is then calculated with these readability indexes.

A. Selection of Readability Parameters

Readability of code is normally linked with comments and naming standards and also called the important factor that impact readability but there are some other aspects the affect the readability. Number of parameters are used in coding that make the code possible and easy to build. There are number of parameters that we find in the literature [2] out of those we have chosen 14 to conduct this comparative study. Table II presents this list of selected parameters.

TABLE II. Parameters Used for Code Readability Comparison

Sr. No.	Parameter	Notation
1	Parenthesis	PAR
2	Indent	IND
3	Spaces	SPA
4	Class Distribution	CD
5	Arithmetic Equations	AE
6	For Loop	FL
7	Nested Loop	NL
8	Do-While Loop	DWL
9	IF-Else	IE
10	Switch	SWI
11	Blanks Lines	BL
12	Line Length (characters)	LL
13	Arrays	ARR
14	Comparison Operators	CO

B. Selection of Code Snippets

A small section of source code or text is called code snippet. Normally they are defined in effective unit of large programs model. In the readability model, first we select the code snippets of Java, C++ and C#. As we know snippets are the small portion of source code, thus we select a small human readable codes that are neither too short nor too long. Each snippet contains a parameter to check their readability impact of that we have discuss earlier. Snippet does not include comments, header functions, and blank lines because they are not meaningful. Secondly code snippets should be logically clear to respondent so, he/she can easily read them. Finally, these snippets are given to the annotators (explain functionality of codes). The ratings for the code snippets are assigned from 1 to 5 where 4 and 5 mean that code is more readable and rank 1 and 2 mean that code is less readable and rank 3 is for average. To perform the online survey, we have used Google Forms and Excel sheets. Respondent can choose one rank (1 to 5) against each language.

V. Comparative Analysis

In this section we perform detailed comparative analysis using the proposed framework presented in the previous section. First we present the details and results of the Survey that we have conducted with the help of programmers of different skill and level.

A. Survey

As mention earlier a set of snippets are selected for human judgment for estimating readability. In Table II, we have presented 14 language constructs that we have chosen to compare the readability of selected programming languages. For every construct we have prepared 6 to 7 pieces of codes for all three languages. Then they are presented to 100 programmers including IT professionals, Programmers, and Computer Science Students. According to their judgment they have ranked snippets. Participant have to rank each snippet from 1 to 5 where 1 is less readable and 5 is more readable.

Each snippet contains a parameter that affects the code readability. And against each parameter participant rank the code readability. Each participant was given the same questionnaires using Google Forms. To improve the visibility results of the survey are presented in bar-chart form in Fig. 3.

We can notice that as per the experts, code snippets written in Java are more readable for almost every selected programming construct. The results also show that C# performs better for two language constructs including DO-While and For Loop is more readable.

B. Code Readability Index

We have computed code readability index for all the selected code snippets against all the selected language constructs using Halstead's metric. Halstead's metric proposed by Maurice Howard Halstead is used to measure the complexity of a program. It depends upon the actual implementation of program which is computed from some operators and operands. It can also computes words size, errors and testing time for C++, C# and Java codes. The parameters used by Halstead's metric are mentioned below:

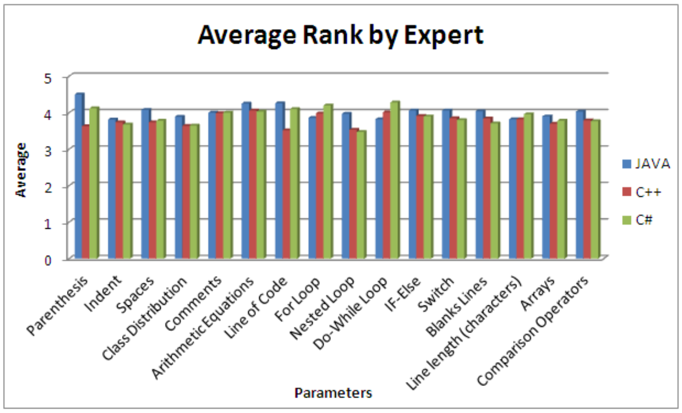


Fig. 3. Results of the Code Readability Comparison

n1: Number of unique operators
n2: Number of unique operands
N1: Total number of operators
N2: Total number of operands

The following list presents the various parameters and their expressions that are offered by Halstead's metric to compute different aspects of programs written in programming languages:

$$Vocabulary : n = n1 + n2$$

$$Size : N = N1 + N2$$

$$Volume : V = length * log2Vocabulary$$

$$Difficulty : D = \left(\frac{n1}{2}\right) * \left(\frac{N1}{n2}\right)$$

$$Efforts : E = Difficulty * Volume$$

$$ProgramLevel : L = V^*/V$$

$$TestingTime : T = \frac{Efforts}{S}, whereS = 18seconds$$

In order to apply the above mentioned metrics on the code snippets, we have developed a source code readability tool (SCRT). SCRT calculates the vocabulary of code, size, volume efforts, errors, testing time and difficulty of the code for all the programs. After calculating these different metrics we have presented the results in upcoming tables including Table III, Table IV, and Table V for Java, C#, and C++, respectively.

After obtaining the results of Halstead's matrices, we have plotted one of the aspects, which is "difficulty" with the help of a line chart to compare the results of all three languages. The results in Fig. 4 clearly show that C++ programs are more difficult to read and understand as compared to the programs written in Java or C#. Mostly Java seems to be less difficult among all the languages in nearly all the language constructs except for comparison operator and arithmetic expressions.

TABLE III. Halstead's Metric Results for Java Language

Params	Vocab	Size	Volume	Difficulty	PRO Level	Quality
PAR	09	21	066.56	03.66	0.46	1.61
IND	19	78	331.33	08.05	0.254	1.037
SPA	8	16	048.00	01.50	0.57	1.115
CD	08	13	039.00	0.125	0.80	NA
CO	09	51	161.66	11.11	0.18	1.12
AE	14	85	323.62	11.25	0.16	1.064
FL	11	27	093.40	03.18	0.42	0.81
NL	13	41	151.71	04.03	0.330	0.60
DWL	13	25	092.51	02.42	0.54	1.76
IE	15	31	012.11	01.83	0.51	0.83
SWI	17	48	196.19	01.41	0.38	1.68
LL	22	43	191.75	03.86	0.51	NA
ARR	14	30	114.22	02.35	0.493	2.07
SCO	13	36	133.21	03.23	0.45	2.1

TABLE IV. Halstead's Metric Results for C# Language

Params	Vocab	Size	Volume	Difficulty	PRO Level	Quality
PAR	08	19	57.0	3.75	0.45	1.41
IND	13	66	244.22	8.46	0.21	0.34
SPA	14	20	76.14	0.53	0.78	1.12
CD	10	17	56.47	0.1	0.73	NA
CO	12	49	175.66	5.83	0.25	1.32
AE	15	73	285.20	7.8	0.20	1.04
FL	11	27	93.41	3.63	0.41	0.85
NL	13	40	148.01	3.76	0.33	0.62
DWL	13	24	88.81	2.15	0.56	1.83
IE	15	26	101.57	1.16	0.61	2.31
SWI	15	42	164.08	1.06	0.38	0.59
LL	20	42	181.52	5.4	0.46	0.97
ARR	14	29	110.41	2.14	0.51	1.89
SCO	11	29	100.32	3.00	0.32	0.52

C. Text Readability Index

Now we calculate the readability of the code with various different text readability indexes. There are many metrics available for the same and among those we have chosen some most popular metrics listed below. After that we have applied them on all the selected code snippets of all three programming languages. The results are presented in Table VI, Table VII, and Table VIII. Before presenting the results, below are the metrics that we have applied to calculate text readability indexes:

- ARI
- FOG
- FKG Level
- SMOG

TABLE V. Halstead's Metric Results for C++ Language

Params	Vocab	Size	Volume	Difficulty	PRO Level	Quality
PAR	11	23	79.56	3.63	0.69	1.54
IND	20	51	220.41	2.97	0.39	0.35
SPA	14	32	121.83	3.21	0.47	0.47
CD	12	50	226.17	5.02	0.47	1.244
COM	09	51	161.66	12.26	0.228	0.34
AE	20	82	354.39	10.5	0.173	0.25
FL	14	36	137.06	4.57	0.40	1.01
NL	17	52	212.54	7.05	0.26	0.31
DWL	15	36	140.64	5.66	0.41	0.549
IE	19	40	169.91	3.36	0.48	0.55
SWI	19	50	212.39	3.13	0.39	0.42
LL	31	101	500.37	8.70	0.26	1.57
ARR	16	41	164	5.06	0.40	0.43
SCO	15	36	140.64	3.96	0.33	0.45

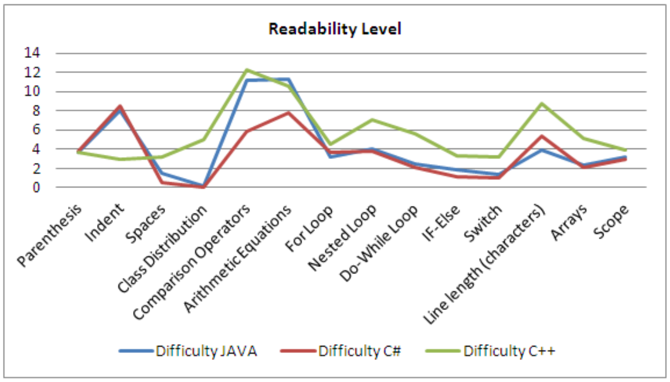


Fig. 4. Comparison of Readability Difficulty among Java, C#, and C++

TABLE VI. Text Readability Index for Java Language

Parameters	ARI	FOG	FKG	SMOG	Average
PAR	11.30	08.41	02.28	08.09	07.52
IND	20.14	11.82	01.73	11.18	11.21
SPA	08.91	12.67	07.49	07.79	09.21
CD	18.90	12.62	11.24	09.24	13.00
AE	01.46	04.68	-00.40	09.00	03.68
FL	15.73	07.60	03.14	13.00	09.86
NL	01.57	05.16	01.50	07.89	04.03
DWL	01.63	06.11	02.61	08.56	04.72
IE	04.11	06.54	02.11	08.09	05.21
SWI	03.39	06.40	01.40	08.83	05.05
SCO	08.39	05.83	01.62	10.14	06.49
LL	06.80	06.15	00.99	08.65	05.64
ARR	18.54	21.85	09.29	11.06	15.18
COM	05.86	05.90	01.66	08.09	05.37

TABLE VII. Text Readability Index for C# Language

Parameters	ARI	FOG	FKG	SMOG	Average
PAR	09.90	03.86	02.75	08.00	06.12
IND	17.38	12.91	02.10	11.30	10.92
SPA	11.88	12.00	07.47	08.00	09.83
CD	21.04	13.30	12.21	09.55	14.03
AE	10.23	06.35	03.31	09.85	07.44
FL	15.05	08.70	05.76	13.63	10.78
NL	04.91	07.02	03.98	08.29	06.05
DWL	05.42	08.65	05.56	08.91	07.13
IE	07.05	08.07	04.4	08.47	06.99
SWI	07.58	07.14	03.71	09.48	06.97
SCO	08.65	07.16	03.87	10.81	07.62
LL	05.97	08.04	03.56	09.08	06.66
ARR	19.05	20.23	09.56	11.24	15.02
COM	06.80	06.15	0.999	08.65	05.64

The obtained results after computing text readability indexes, are plotted with the help of bar-chart. Fig. 5 shows the results for all three programming languages against all programming constructs. The results again show that Java language codes are more readable as compare to C# and C++. But in some constructs such as comparison operators, arithmetic equations and scope C# is more readable as per text readability index.

VI. Statistical Analysis

In this section we present statistical analysis that we have performed on the results obtained after experiments. For this we have chosen *T-test* for the same. The *T-test* is used to compare the two sample means. Where one sample means

TABLE VIII. Text Readability Index for C++ Language

Parameters	ARI	FOG	FKG	SMOG	Average
PAR	8.38	04.66	01.69	07.35	05.52
IND	09.3	11.91	08.29	09.48	09.74
SPA	03.08	06.54	01.51	07.79	04.73
CD	15.09	13.02	09.90	11.00	12.25
AE	02.94	06.22	02.05	08.09	04.82
FL	19.72	18.83	13.98	11.94	16.12
NL	02.69	08.65	03.99	07.69	05.75
DWL	05.38	10.80	06.43	08.47	07.77
IE	03.19	08.14	04.16	08.09	05.89
SWI	04.38	08.82	05.30	08.65	06.79
SCO	09.3	11.20	08.20	09.4	09.53
LL	-00.39	07.20	03.36	07.35	04.38
ARR	09.88	10.31	05.06	12.27	09.38
COM	04.59	08.66	03.6	08.00	06.22

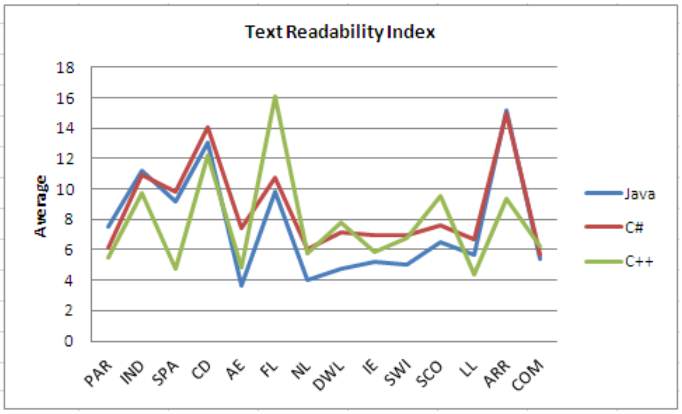


Fig. 5. Comparison of Text Readability Index among Java, C#, and C++

can be paired with other sample mean observation. In paired T-Test each entity is measured twice, result will be given in pairs. Table IX presents Halstead's arithmetic mean, standard deviation and standard mean error are given for all three languages (Java, C#, C++). Table X presents paired correlation between Java and Halstead index, C# and Halstead index and C++ and Halstead index of C++. Where correlation r ¿ 0.50 shows the strong relationship and r ¡ 0.50 shows the weak positive relationship.

TABLE IX. Mean, Std. Deviation, and Std. Error Mean

	Mean	N	Std. Deviation	Std. Error
Java	4.0089	99	0.27024	0.02716
Halstead Java	3.3005	99	0.81316	0.08173
C#	3.8054	99	0.51242	0.05150
Halstead C#	3.7690	99	0.82096	0.08251
C++	3.8856	99	0.32749	0.03291
Halstead C++	3.8871	99	0.41251	.04146

TABLE X. Correlation between Mean, Std. Deviation, and Std. Error Mean

	N	Correlation	Sig.
Java & Metric Readability-Java	99	0.730	0.191
C# & Metric Readability-C#	99	0.529	0.001
C++ & Metric Readability-C++	99	0.610	0.553

The statistical results show that Java programming language has been found being more readable as compared to other programming languages.

VII. Conclusion

Code readability influences maintenance of a software at great deal. Due to its salient importance, we have conducted a comparative study to estimate readability of the codes produced by Java, C#, and C++ programming languages. We identify important language constructs that affect the code readability and then propose a novel framework to compare the codes using three different dimensions. First we have performed an expert survey involving programmers and experts to judge the readability of codes. Then we have applied code readability and text readability indexes to again calculate readability of the same programs. We have computed these indexes using a source code readability tool (SCRT). The experiment results show that Java language produces more readable code as compared to C# and C++. Only for a few language constructs like comparison operators and arithmetic operator. We have also statistically analyzed the results using SPSS tool to verify the effectiveness of experiments. This analysis also verifies that Java language code is more readable than C# and C++.

VIII. Future Work

In future we are planning to extend our analysis on other famous languages also including Python and VB.NET. Other than these, we are also looking to conduct an analysis on programming languages that are used specifically for mobile application development.

References

[1] S. Fakhoury, D. Roy, A. Hassan and V. Arnaoudova, ”Improving Source Code Readability: Theory and Practice,” 2019 IEEE/ACM 27th International Conference on Program Comprehension (ICPC), pp. 2-12, Montreal, QC, Canada, 2019.

[2] Buse, R.PL., Westley R.W. ”A metric for software readability.” Proceedings of the 2008 international symposium on Software testing and analysis, pp. 121-130, Seattle, WA, USA, July 20-24, 2008.

[3] Pahal, Ankit, and Rajender S. Chillar. ”Code Readability: A Review of Metrics for Software Quality.”International Journal of Computer Trends and Technology (IJCTT) – Volume 46 Number 1- April 2017

[4] Deepa D., Dua A. K. “Evaluation of Quality of Source Code By Code Readability. International Journal of Advanced Research in Computer Science and Software Engineering, 2015.

[5] Tashtoush, Y. ”Impact of programming features on code readability.” International Journal of Software Engineering and its Applications. 7(6):441-458. November 2013.

[6] Relf, P.A., ”Tool assisted identifier naming for improved software readability: an empirical study”, In International Symposium on Empirical Software Engineering, Noosa Heads, Qld., Australia, November 17-18, 2005.

[7] Sivaprakasam., P, Sangeetha., V. “Improving software qualitythrough the development ofcode readability” International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 6, August 2012.

[8] Relf, P.A., ”Achieving software quality through source code readability”, Quality Contract Manufacturing LLC., 2004.

[9] DeYoung, G.E., Kampen, G.R. and Topolski, J.M. ”Analyzer-generated and human-judged predictors of computer program readability.” Proceedings of the 1982 conference on Human factors in computing systems. pp 223-228, Gaithersburg, Maryland, USA, March 15-17, 1982.

[10] Sivaprakasam, P., and V. Sangeetha. ”An accurate model of software code readability.” International Journal of Engineering Research and Technology.ESRSA Publications (2012).

[11] Collar Jr, Emilio, and Ricardo Valerdi. ”Role of software readability on software development cost.” 2006.

[12] Aggarwal, Krishan K., Yogesh Singh, and Jitender Kumar Chhabra. ”An integrated measure of software maintainability.” Reliability and maintainability symposium, 2002.Proceedings.Annual.IEEE, 2002.

[13] Daryl, P., Hindle, A., and Devanbu, P., ”A simpler model of software readability”, In Proceedings of the 8th working conference on mining software repositories, pp. 73-82, Waikiki, Honolulu, HI, USA, May 21-22, 2011.

[14] McLaughlin, G. Harry. ”SMOG grading-a new readability formula.” Journal of reading 12.8 (1969): 639-646.

[15] TIOBE. https://www.tiobe.com/tiobe-index/. (accessed on October 17, 2019).

[16] Automated Readability Index (ARI). https://en.wikipedia.org/wiki/Automated_readability_index. (accessed on October 18, 2019).

[17] Kincaid, J.P., Fishburne, R.P., Rogers, R.L., & Chissom, B.S., ”Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel.” Research Branch Report 8–75. Chief of Naval Technical Training: Naval Air Station Memphis. 1975.

[18] Gunning, Robert., ”The Technique of Clear Writing”. McGraw-Hill. pp. 36–37. 1952.

[19] Senter, R.J.; Smith, E.A. (November 1967). ”Automated Readability Index”. Wright-Patterson Air Force Base: iii. AMRL-TR-6620. Retrieved March 18, 2012.

CA-PCS: A Cellular Automata based Partition Ciphering System

Fatima Ezzahra Ziani[1], Anas Sadak[2], Charifa Hanin[3], Bouchra Echandouri[4], Fouzia Omary[5]
Faculty of Sciences, University Mohammed V
Computer Science Department
Rabat, Morocco

***Abstract*—In this paper, the authors present a modified version of the Partition Ciphering System (PCS) encryption system previously proposed. The previously developed encryption system PCS uses the partition problem to encrypt a message. The goals of newly developed system are avoiding statistical and frequency attacks, by providing a balance between 0s and 1s, ensuring a good level of entropy and achieving confidentiality through encryption. One of the novelties of the new design compared to its predecessor is the use of cellular automata (CAs) during the encryption. The use of CAs is justified by their good cryptographic properties that provide a level of security against attacks, and better confusion and diffusion properties. The new design is first presented with details of the encryption and decryption mechanisms. Then, the results of the DIEHARDER battery of tests, the results of the avalanche test, a security analysis and the performance of the system are outlined. Finally, a comparison between CA-PCS and PCS as well as the AES encryption system is provided. The paper shows that the modified version of PCS displays a better performance as well as a good level of security against attacks.**

Keywords*—*Partition ciphering system; partition problem; frequency analysis; cellular automata; avalanche effect; confusion; diffusion; statistical properties; cryptographic properties

I. Introduction

One of the five pillars of cryptography is achieving confidentiality. This latter comprises two principles: data confidentiality and privacy. Data confidentiality ensures that no data is accessed or revealed to unauthorized parties. Privacy controls the access to data and storage of data by concerned parties [1]. This paper presents a modified version of the Partition Ciphering System (PCS), which was previously developed by the authors [2]. It is a symmetrical encryption system based on the partition problem, more precisely the Card-Partition version. The use of the partition problem in PCS was motivated by the fact that it changes the frequency of the appearance of characters between the plaintext and the ciphertext. Consequently, PCS is robust against frequency cryptanalysis; an adversary cannot learn any information about the plaintext from the ciphertext. However, PCS has some limitations to check the diffusion property and resistance to some attacks like linear and differential attacks. A cellular automaton (CA) is a suitable candidate to provide better confusion and diffusion. Also, the CA cryptographic properties could be studied to verify the security level. These later are nonlinearity, algebraic degree, balancedness, resiliency, and correlation immunity. A CA is a dynamic system involving a network of cells. CAs are widely used in cryptography and other fields to benefit from their simplicity, parallelism, and unpredictability. Besides, CAs make the hardware and software implementations easier [3]. In this paper, a new design called CA-PCS (Cellular Automata based Partition Ciphering System) is proposed. It consists of a hybrid CA, with satisfying cryptographic properties, that evolves multiple iterations to increase resistance to linear and differential attacks, followed by the insertions of necessary blocks so that the frequency of all the blocks is the same. In addition to a random permutation is applied to the results of the second step. Each layer produces better confusion and diffusion, and consequently, better resistance to linear and differential cryptanalysis. Also, the cryptographic properties of the CA ruleset are studied and display good results. A high nonlinearity, high algebraic degree, and balancedness are satisfied. CA-PCS was compared to AES and PCS in terms of randomness, security, and performance. Thus, the CA-PCS results are satisfying.

The rest of this article is organized as follows: In Section 2, a brief background on cellular automata is presented. Next, in Section 3, the related works are included. Then, CA-PCS is detailed in Section 4. Section 5 provides a brief description of the PCS and AES encryption systems. Finally, Section 6 presents results and security analysis.

II. Background on Cellular Automata

The history of cellular automata goes back to the 1940s when Stanislaw Ulam [4] initiated their study by taking interest in self-replicating automata. Then in the 1960s, John von Neumann used them in Biology for modeling self-reproduction [5]. They were later on popularized by John Conway's game of life in the 1970s [6]. They were first use in cryptography by Stephen Wolfram in the 1980s [7]. Simply put, a cellular automaton is a network of cells, each of which has a state that changes from a time step t to a time step t+1 according to a defined local rule and depending on its neighbors. The interest of the scientific community in cellular automata stems from the fact that simple local calculations at the cells scale produce a complex behavior at the automaton scale. Another interesting aspect of using cellular automata is that both uniformity and non-uniformity can be modeled. A cellular automaton is defined as [3] (d, L, S, N, f), where d represents the cellular space dimension, L represents the cellular space, S is the finite set of states, N is the neighborhood vector and f or (f_1,f_2,...) is the local rule or ruleset respectively. The global rule of the cellular automata is designated by Φ.

By modifying the tuple (d, L, S, N, f), different kinds of cellular automata can be obtained. One interesting type of cellular automata was introduced by Wolfram in [8]. This kind

TABLE I. An example of a linear and nonlinear ECA rule

Rule	105		135	
Linear?	Yes		No	
Algebraic Normal Form	$1 \oplus x_{i-1} \oplus x_i \oplus x_{i+1}$		$1 \oplus x_i.x_{i+1} \oplus x_{i-1}$	
Truth table	111	0	111	1
	110	1	110	0
	101	1	101	0
	100	0	100	0
	011	1	011	0
	010	0	010	1
	001	0	001	1
	000	1	000	1

of CAs is called Elementary Cellular Automata (ECAs). They are one-dimensional, two-state (0 or 1), 3-neighborhood CAs. They are of particular interest in cryptography as their simple implementation, both in hardware and software, their good cryptographic properties and the small number of possible rules ($2^{2^3} = 256$) are well suited in this field as they can be thoroughly studied. The local rules can be either linear (only XOR operator $\oplus$ in their Boolean expression) or nonlinear (AND($\cdot$)/OR($+$) operators as well in their Boolean expression). Table I shows an example of a linear and nonlinear rule.

III. Related Work

The partition problem or Equal Piles Problem, which is the source of inspiration for this work, was first studied by Jones and Beltramo in [9], where they defined a challenging instance. They tried nine standard genetic algorithms, but without finding an optimal solution. To solve this instance of the problem, Falkenauer [10] and William [11] proposed particular types of genetic algorithms. Concretely, Falkenauer [10] tried to adjust the grouping genetic algorithm that he designed previously using specific crossover and mutation operators suited for similar problems. William [11]used a particular approach in the design of the Eager Breeder genetic algorithm, which makes the manipulation of genetic materials easier and produces better results compared to the previous algorithms. However, their results are not that good for this article's proposed design.More recently, evolutionist algorithms were also used to come up with a solution to the partition problem as in the works of Trichni [12], Bougrine [13] and Kaddouri [14].

The first use of cellular automata in cryptography goes back to Wolfram in [7]. He applied rule 30 to design a pseudorandom number generator (PRNG) and a stream cipher. A more recent example of the use of CAs in an encryption algorithm is the design of Das et al. [15] who proposed a block cipher using one dimensional programmable CAs. Other works using one dimensional uniform CAs include Bhaumik [16] and Roy [17]. Non uniform one-dimensional CAs were studied by Mehta [18] and Bouchkaren [19]. Two dimensional uniform CAs were used by Bouchkaren [20] and Faraoun [21]. CAs were also used for image encryption by Li in [22], who made use of two dimensional non-uniform CAs. Other image encryption schemes can be found in [23] and [24].

IV. CA-PCS Design

A. CA-PCS Encryption Algorithm

The CA-PCS encryption scheme goes through three steps:

1) CA Evolution: The first step includes the hybrid CA evolving of the binary message using the rules $\{90, 150, 30, 180, 45, 90, 150, 30\}$. Linear rules 90 and 150 provide better diffusion property and high cycle length [25]. While nonlinear rules 30, 45, and 180 provide better confusion property [26]. Moreover, these rules provide resistance to linear attacks and differential attacks. Because of the high nonlinearity met after a few iterations and the significant algebraic degree.

2) Blocks Insertion: The second step consists of representing the first step's result as a partition and add some blocks at random positions to get the same appearance frequency for all blocks. At first, the CA output is split into blocks of a randomly chosen size $2 \leq k \leq 16$. Then the ideal cardinality IC is computed $IC = max\{Card(L_1), Card(L_2), ..., Card(L_m)\}$. Next, for each block B_i, the cardinality of the corresponding L_i, representing the positions of Bi in the CA output, is compared to the IC. Accordingly, if $Card(L_i) < IC$, then B_i is inserted in a random position $1 < P_{ij} < size(CAoutput)$ where $0 < j < IC - Card(L_i)$. Next, the P_{ij} is inserted in the ListOfInsertedBlocksPositions.

3) Permutation: Finally, a random permutation is applied to the set $\{L_1, L_2, ..., L_m\}$. This permutation is useful to change the blocks' occurrence lists L_is. It is denoted formally by $\pi : S \rightarrow S$ where S is a set of m elements. m! permutation of $\{L_1, L_2, ..., L_m\}$ are possible. A possible example of a random permutation for m=10, π :{L_1, L_2, L_3, L_4, L_5, L_6, L_7, L_8, L_9, L_{10}} $\rightarrow$ $\{L_2, L_4, L_1, L_6, L_3, L_9, L_7, L_{10}, L_8, L_5\}$. Following this example, $L_1 \rightarrow L_2$, $L_2 \rightarrow L_4$, $L_3 \rightarrow L_1$, $L_4 \rightarrow L_6$, $L_5 \rightarrow L_3$, $L_6 \rightarrow L_9$, $L_7 \rightarrow L_7$, $L_8 \rightarrow L_{10}$, $L_9 \rightarrow L_8$, $L_{10} \rightarrow L_5$. Accordingly, B_1 will appear in the positions of B_2, B_2 will appear in those of B_4, and so on.

4) Key generation: The secret key comprises four elements:
$SK = \{k, CASeq, ListOfInsertedBlocksPositions, PSeq\}$
The random integer k is the blocks size. The CASeq binary sequence where $CASeq = M \oplus M'$ where M is the plaintext, and M' is the output of the CA evolution step. The ListOfInsertedBlocksPositions which comprises the positions where blocks are inserted. the PSeq binary sequence $PSeq = M" \oplus C$ where M" is the output of the blocks insertion step and C is the ciphertext. Fig. 1 summarizes the encryption process of CA-PCS.

B. CA-PCS Decryption Algorithm

The CA-PCS decryption process, as Fig. 2 displays, is as follows, given the ciphertext C and the secret key
$SK = \{k, CASeq, ListOfInsertedBlocksPositions, PSeq\}$:
At first, the PSeq sequence is XORed with the ciphertext to get M". Then, M" is split into blocks of size k. Next, inserted blocks are removed from M" using the ListOfInsertedBlocksPositions to get M'. Then M' is XORed with the CASeq to get the plaintext.

Algorithm 1 CA-PCS Encryption Algorithm

```
Input: The message M
Output: The ciphertext C and the secret key K
Begin
$it \leftarrow 64$
$size \leftarrow sizeOf(M)$
$k \leftarrow random(2, 16)$                          ▷ random integer2 $< k \leq 16$
$ruleSet \leftarrow \{30, 90, 150, 30, 180, 45, 90, 150\}$
for $0 < j \leq it$ do
    for $0 < i \leq size$ do
        $x \leftarrow i - 1 \bmod sizeOf(ruleSet)$
        if $(x == 0)||(x == 3)$ then                                ▷ 30
            $M'[i] \leftarrow M[i-1] \oplus (M[i] + M[i+1])$
        else if $(x == 1)||(x == 6)$ then                           ▷ 90
            $M'[i] \leftarrow M[i-1] \oplus M[i+1]$
        else if $(x == 2)||(x == 7)$ then                          ▷ 150
            $M'[i] \leftarrow M[i-1] \oplus M[i] \oplus M[i+1]$
        else if $x == 4$ then                                      ▷ 180
            $M'[i] \leftarrow M[i-1] \oplus (M[i].(1 \oplus M[i+1]))$
        else                                                        ▷ 45
            $M'[i] \leftarrow M[i-1] \oplus (M[i] + (1 \oplus M[i+1]))$
        end if
    end for
end for
$CASeq \leftarrow M \oplus M'$
$M" \leftarrow DivideIntoBlocks(M', k)$
$n \leftarrow sizeOf(M")$
$m \leftarrow NumberOfDifferentBlocks(M")$
$Partition \leftarrow ToPartition(M")$
$ListOfBlocks \leftarrow DifferentBlocks(M")$          ▷ $\{B_1, ..., B_m\}$
$IC \leftarrow ComputeIdealCardinality(PlaintextPartition)$
for $1 \leq i \leq m$ do
    while $Card(L_i) < IC$ do
        $M" \leftarrow insert(B_i, M", randomPosition)$
        $Insert(ListOfInsertedBlocksPositions, randomPosition)$
        $Insert(L_i, randomPosition)$
    end while
end for
$permutation \leftarrow generateRandomPermutation(\{1, 2, ..., m\})$
$Ciphertext \leftarrow applyPermutation(M", permutation)$
$PSec \leftarrow Ciphertext \oplus M"$
$secretK \leftarrow \{k, CASeq, ListOfInsertedBlocksPositions, PSec\}$
End
```

Algorithm 2 Decryption algorithm

```
Input:The secret key SK and the ciphertext C
Output: The message M
Begin
$M" \leftarrow C \oplus PSec$
$M^{(3)} \leftarrow DivideIntoBlocks(M", k)$
for  $i$ from $sizeOf(ListOfInsertedBlocksPositions)$ to 1  do
    $M' \leftarrow Remove(M^{(3)}, ListOfInsertedBlocksPositions[i])$
end for
$M \leftarrow M' \oplus CASeq$
End
```

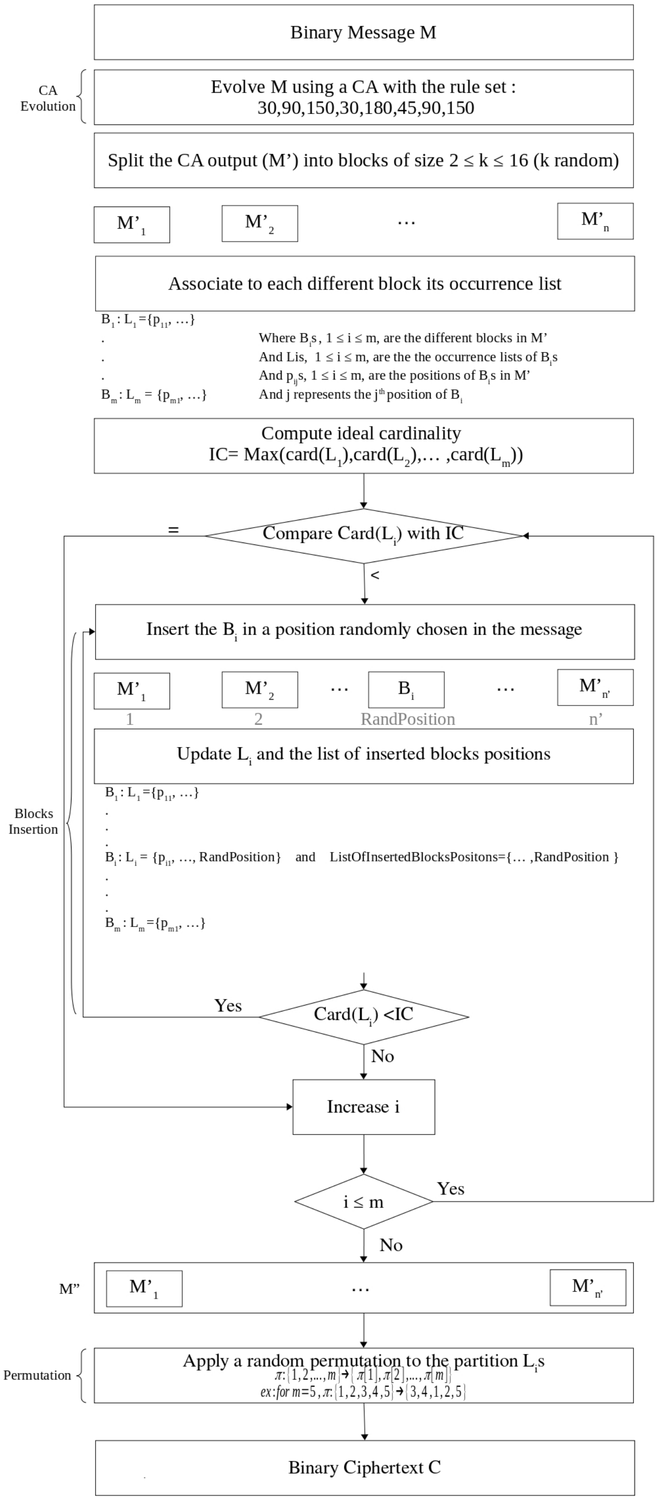

Binary Message M
CA Evolution
Evolve M using a CA with the rule set :
30,90,150,30,180,45,90,150
Split the CA output (M') into blocks of size 2 ≤ k ≤ 16 (k random)
M'1
M'2
…
M'n
Associate to each different block its occurrence list
B1 : L1 ={p11, …}
Bm : Lm = {pm1, …}
Where Bis , 1 ≤ i ≤ m, are the different blocks in M'
And Lis, 1 ≤ i ≤ m, are the the occurrence lists of Bis
And pijs, 1 ≤ i ≤ m, are the positions of Bis in M'
And j represents the jth position of Bi
Compute ideal cardinality
IC= Max(card(L1),card(L2),… ,card(Lm))
=
Compare Card(Li) with IC
<
Insert the Bi in a position randomly chosen in the message
M'1
M'2
…
Bi
…
M'n'
1
2
RandPosition
n'
Update Li and the list of inserted blocks positions
B1 : L1 ={p11, …}
Bi : Li = {pi1, …, RandPosition}   and   ListOfInsertedBlocksPositons={… ,RandPosition }
Bm : Lm ={pm1, …}
Blocks Insertion
Yes
Card(Li) <IC
No
Increase i
Yes
i ≤ m
No
M"
M'1
…
M'n'
Permutation
Apply a random permutation to the partition Lis
π:{1,2,...,m}→{π[1],π[2],...,π[m]}
ex:for m=5, π:{1,2,3,4,5}→{3,4,1,2,5}
Binary Ciphertext C

Fig. 1. CA-PCS Encryption

V. The PCS and AES description

This section presents a brief description of a previously developed scheme Partition Ciphering System (PCS) and the Advanced Encryption Standard (AES).

A. Partition Ciphering System (PCS)

The Partition Ciphering System PCS [2] is a symmetric enryption schemme that encrypts a plaintext in three steps. the first step consists of the construction of a partition from the plaintext, which is initially split into blocks of size k>2. Each block is associated with a list of occurrences. This partition undergoes some transformations in a way to make the ciphertext resistant to frequency cryptanalysis. Next, the ideal cardinality IC is computed : let $c = \frac{n}{m}$, where n is the number of blocks in the plaintext, and m is the number of different blocks in

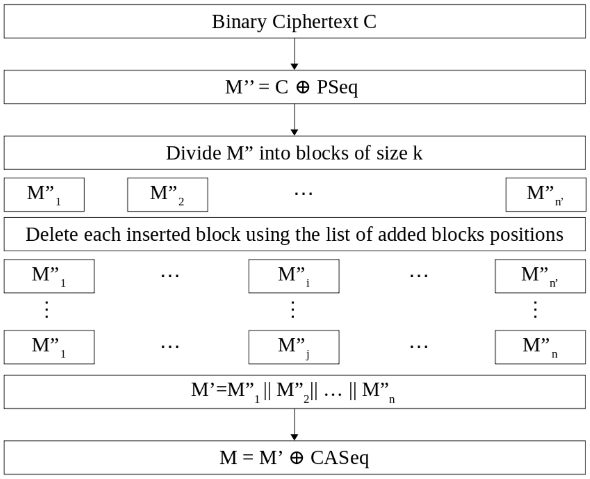

Binary Ciphertext C
M'' = C ⊕ PSeq
Divide M'' into blocks of size k
M''₁
M''₂
…
M''ₙ'
Delete each inserted block using the list of added blocks positions
M''₁
…
M''ᵢ
…
M''ₙ'
M''₁
…
M''ⱼ
…
M''ₙ
M'=M''₁ || M''₂|| … || M''ₙ
M = M' ⊕ CASeq

Fig. 2. CA-PCS Decryption

the plaintext. If $c \in \mathbf{N}$ then $IC = c$ else $IC = \lceil c \rceil$. This cardinality defines the number of occurrences of each block in the ciphertext. In the last step, the blocks B_is are inserted or deleted according to the cardinal of the corresponding list of appearances L_i. When the $Card(L_i) < IC$, then the block B_i is appended to the message. When the $Card(L_i) > IC$, then the block B_i is deleted from a random position.

B. Advanced Encryption Standard (AES)

Advanced Encryption Standard (AES) [27] is a symmetric cipher that encrypts 128-bit blocks using keys of size 128 bit, 192 bit, or 256 bit. It comprises N rounds, where N changes according to the length of the key: 10 for a 128-bit key, 12 for a 256-bit key, and 14 for a 192-bit key. In the first step, the plaintext is XORed by the first 128 bit of the key. Next, for N-1 iteration, four operations are performed: SubBytes, ShiftRows, MixColumns, and AddRoundKey. [27] provides a detailed description of these operations. Finally, the last round consists of only SubBytes, ShiftRows, and AddRoundKey operations.

VI. Results and Security Analysis

This section displays the statistical tests and the confusion and diffusion properties of CA-PCS compared to the AES.

A. Dieharder Test

The battery of tests Dieharder was designed by Robert G. Brown to check out the behavior of PRNGs and cryptographic primitives like encryption systems, hash functions, and MACs. It involves tests from diehard, some NIST tests, and other tests developed by Brown and Bauer [28]. The authors generated three files of 10 Mb using PCS, CA-PCS, and AES ciphers. Then, they run the battery overs these files. Table II displays the results. The P-values are the probability that the generated sequences are random. If 0.005<P-value<0.995, then the systems pass the test. Since 0.10<P-values(PCS)<0.91, 0.2<P-values(CA-PCS)<0.92, and 0.005<P-values(AES)<0.95, then all the systems pass all the tests. Also, the P-values of the ciphers are uniformly distributed in the range [0, 1], to conclude, CA-PCS displays good results regarding the statistical tests compared to PCS and AES.

TABLE II. Dieharder Results of CA-PCS, AES, and PCS

Tests names	P-values CA-PCS	P-values AES	P-values PCS
Diehard birthdays	0.5357	0.0836	0.8625
Diehard operm5	0.4946	0.0967	0.8971
Diehard rank 32x32	0.5887	0.7711	0.1402
Diehard rank 6x8	0.7192	0.6936	0.3240
Diehard bitstream	0.4615	0.6593	0.4530
Diehard opso	0.5559	0.7204	0.3559
Diehard oqso	0.5092	0.6363	0.1898
Diehard dna	0.5686	0.3142	0.2811
Diehard count 1s str	0.4114	0.8797	0.8988
Diehard count 1s byt	0.6995	0.8451	0.7611
Diehard parking lot	0.2622	0.8514	0.773
Diehard 2dsphere	0.4555	0.5370	0.7910
Diehard 3dsphere	0.6735	0.3863	0.2487
Diehard squeeze	0.6888	0.8732	0.7991
Diehard sums	0.9130	0.0058	0.1779
Diehard runs	0.2342	0.3810	0.7702
Diehard craps	0.7063	0.8630	0.9093
Marsaglia tsang gcd	0.6682	0.7107	0.4046
Sts monobit	0.5815	0.6915	0.54319
Sts runs	0.4394	0.4656	0.1070
Sts serial	0.6616	0.5643	0.6388
Rgb bitdist	0.6689	0.5724	0.4844
Rgb minimum distance	0.5515	0.3475	0.4441
Rgb permutations	0.6639	0.6588	0.4145
Rgb lagged sum	0.5074	0.5363	0.6067
Rgb kstest test	0.2840	0.4934	0.1025
dab bytedistrib	0.5920	0.4758	0.2636
dab dct	0.8842	0.9448	0.8735
dab filltree	0.4757	0.4721	0.5212
dab filltree2	0.8987	0.7090	0.3727
dab monobit2	0.8994	0.0507	0.6055

B. Confusion and Diffusion Tests

This section presents the confusion and diffusion properties of the CA-PCS system in comparison with AES. A secure encryption system from statistical analysis, as stated by Shannon [29], has good confusion and diffusion properties (e.g., AES is a secure system). If the relation between the ciphertext and the secret key is hidden, then the confusion property is verified. In other terms, replacing one bit in the secret key has an impact on most of the bits in the ciphertext. If the relation between the plaintext and the ciphertext is masked, then the diffusion property is checked. In other words, changing one bit in the plaintext affect almost all the bits of the ciphertext. Fig. 3 shows the confusion property for CA-PCS compared to the AES. According to Fig. 3, the percentage of the changed bits in the ciphertext is approximately 50% for CA-PCS and AES. Concretely, the values for CA-PCS are between 0.40% and 0.61%, while the values for AES are between 0.36% and 0.61%. These values confirm that CA-PCS has better confusion property. Fig. 4 illustrates the diffusion property of CA-PCS and AES. The mean value of the percentages of changed bits in the ciphertext is nearly 50%. The values for CA-PCS are

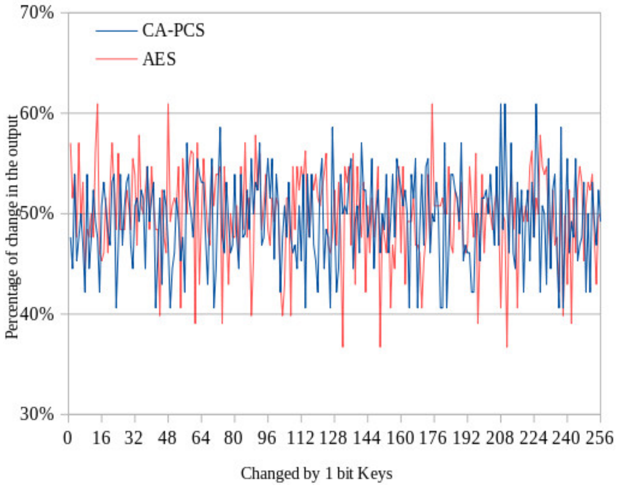


Fig. 3. Confusion Test of CA-PCS and AES

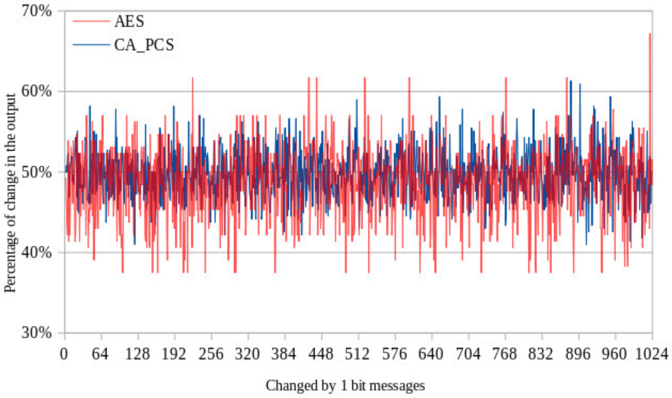


Fig. 4. Diffusion Test of CA-PCS and AES

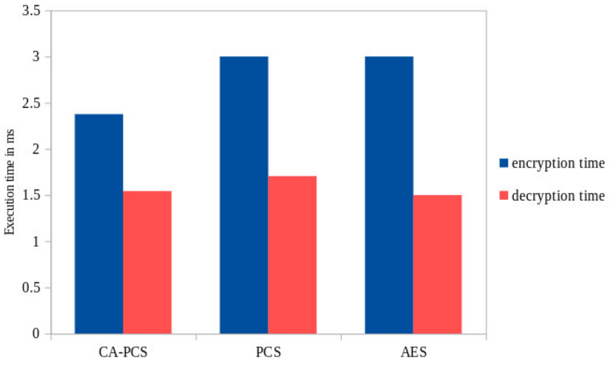


Fig. 5. Encryption and decryption time of CA-PCS, PCS and AES

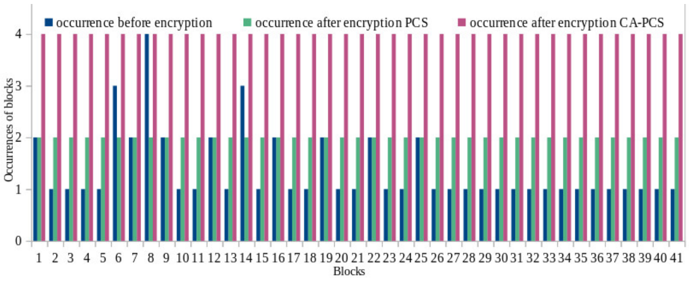


Fig. 6. Frequency of blocks before and after encryption for CA-PCS and PCS

TABLE III. NONLINEARITY

Iterations	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	2	0	0	2	2	2	0	0
2	8	8	8	8	8	8	8	8
3	32	48	48	48	28	44	48	48

between 41% and 61%, and the values for AES are between 37% and 67%. Consequently, CA-PCS has better diffusion.

C. Encryption and Decryption Time of CA-PCS, AES and PCS

This part (Fig. 5) compares the encryption and decryption time of CA-PCS with the previously developed scheme PCS and AES. Fig. 5 shows that CA-PCS requires less time in the encryption process compared to PCS and AES. While the PCS and AES take the same time to encrypt. The time of decryption is approximately the same for CA-PCS, PCS, and AES. To conclude CA-PCS displays good results.

D. Frequency Analysis

This part presents the frequency analysis of the outputs of CA-PCS and PCS. As mentioned in [2], the purpose was to have a ciphertext with blocks appearing with the same frequency, so that frequency analysis does not reveal any information about the plaintext. As CA-PCS is an improved version of PCS, the same objective persists. CA-PCS is different from PCS in all steps. The CA evolution is the first step of CA-PCS. Next, the ideal cardinality computation. Later, the insertion of blocks follows. The resulting intermediate output undergoes a permutation. While in PCS, the ideal cardinality is computed in a way to have blocks to add or remove. The objective of CA-PCS design is to provide better confusion and diffusion, in addition to resistance to some attacks like linear and differential attacks. Fig. 6 represents the frequency analysis performed on the outputs of CA-PCS and PCS for the same plaintext. Fig. 6 shows that frequency analysis will never divulge any information. As a result, frequency cryptanalysis is impossible.

E. Cryptographic Properties of the Ruleset Used in the CA Evolution

This section presents the cryptographic properties, namely, nonlinearity, algebraic degree, correlation immunity, resiliency, and balancedness, of the CA ruleset $\{30, 90, 150, 30, 180, 45, 90, 150\}$. It is applied alternately on the CA cells in the evolution step. In this section, to study the ruleset, an example of 8 cells is considered. Tables III to VII shows the variation of the cryptographic properties with iterations.

Nonlinearity and algebraic degree increase significantly within iterations. Also, balancedness persists. The resiliency and the correlation immunity decrease with iterations because high nonlinearity affects the level of resiliency and correlation immunity. Most of the cryptographic systems require high

TABLE IV. ALGEBRAIC DEGREE

Iterations	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	2	1	1	2	2	2	1	1
2	3	2	2	3	3	3	2	2
3	4	3	3	4	5	4	3	3

TABLE V. RESILIENCY

Iterations	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	0	1	2	0	0	0	1	2
2	0	2	2	0	1	0	2	2
3	0	0	0	0	0	0	0	1

TABLE VI. CORRELATION IMMUNITY

Iterations	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	0	1	2	0	0	0	1	2
2	0	2	2	0	1	0	2	2
3	0	0	0	0	0	0	0	1

TABLE VII. BALANCEDNESS

Iterations	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	✓	✓	✓	✓	✓	✓	✓	✓
2	✓	✓	✓	✓	✓	✓	✓	✓
3	✓	✓	✓	✓	✓	✓	✓	✓

TABLE VIII. BRUTE FORCE ATTACK OF AES, PCS AND CA-PCS

Encryption schemes	AES-128	AES-192	AES-256	PCS	CA-PCS
Key length	128 bit	192 bit	256 bit	$\geq$ 256 bit	$\geq$ 256 bit
# possible keys	2^{128}	2^{192}	2^{256}	$\geq 2^{256}$	$\geq 2^{256}$
Security level	near term	near term	long term	long term	long term

nonlinearity, and algebraic degree as well as balancedness. These cryptographic properties are important to avoid attacks, particularly linear attacks, differential attacks, and statistical cryptanalysis.

F. Brute-Force Attack

In a brute-force attack, the attacker tests each possible key to get a comprehensible plaintext from the transformation of the ciphertext [1]. The key length is considered the security parameter that provides the security level of the studied system. This attack needs more time and resources to get the right key when the key length is high. It can be impossible unless an attacker has a quantum computer. If the level of security desired is for the near term, then a symmetric key of at least 128 bit is used. The key should be of at least 256 bit to reach long term security. Since the AES has three versions, AES-128, AES-192, and AES-256, both security levels can be satisfied. PCS, from [2], has a secret key of size greater than 256 bit. Also, CA-PCS has a secret key of at least 256 bit. Unless an attacker has a quantum computer, he cannot get the secret key to decrypt to an intelligible plaintext. Table VIII summarises the security level of AES, PCS, and CA-PCS.

G. Linear and Differential Attacks

Linear attack analyzes the linear approximations of the plaintext, the ciphertext, and the secret key [30]. It is a known-plaintext attack, while differential attack studies the differences between plaintexts and ciphertexts [31]. It is a chosen-plaintext attack. A cipher should be robust against the linear and differential attacks. The confusion property, which is satisfied using the nonlinear parts of the system, is necessary to resist these types of attacks. In general, S-Boxes are responsible for this purpose. But, other primitives, like nonlinear cellular automata, can lead to the same results. In CA-PCS, the ruleset used to evolve the CA has high nonlinearity, and maintain the balancedness. These features make these attacks difficult for a cryptanalyst.

VII. CONCLUSION

In this article, an enhanced version of PCS, a previously developed encryption scheme, is proposed. The proposed system, called CA-PCS, makes use of cellular automata to increase the security level of the design. Precisely, the ruleset used provides satisfying results in terms of cryptographic properties, randomness tests, confusion, and diffusion properties. Linear and differential attacks are difficult to achieve because of the high non-linearity and the high algebraic degree provided by the ruleset. Also, the balancedness and the randomness produce resistance to statistical cryptanalysis. Moreover, CA-PCS is robust against brute force attacks. Besides, the performance of CA-PCS is better than PCS and AES. In future work, the authors will extend the proposed scheme to ensure authentication.

REFERENCES

[1] W. Stallings, Cryptography and network security: principles and practice. Pearson Prentice Hall, 2017.

[2] F. E. Ziani and F. Omary, "Partition Ciphering System: A Difficult Problem Based Encryption Scheme," International Journal of Advanced Computer Science and Applications, vol. 10, no. 11, 2019

[3] K. Bhattacharjee, N. Naskar, S. Roy, and S. Das, "A survey of cellular automata: types, dynamics, non-uniformity and applications," Natural Computing, 2018.

[4] S. Ulam, "Random processes and transformations," in Proceedings of the International Congress on Mathematics, vol. 2, pp. 264-275, 1952.

[5] J. T. Schwartz, J. V. Neumann, and A. W. Burks, "Theory of Self-Reproducing Automata," Mathematics of Computation, vol. 21, no. 100, p. 745, 1967.

[6] M. Gardner, "On cellular automata self-reproduction, the garden of eden and the game of Life," Scientific American, vol. 224, no. 2, pp. 112 - 118, 1971.

[7] S. Wolfram, "Cryptography with Cellular Automata," Lecture Notes in Computer Science Advances in Cryptology - CRYPTO-85 Proceedings, pp. 429 - 432.

[8] S. Wolfram, A new kind of science. Champaign, IL: Wolfram Media, 2002.

[9] D. R. Jones and M. A. Beltramo, " Solving Partitionning Problems with Genetic Algorithms," Proceedings of the Fourth International Conference on Genetic Algorithms, 1991.

[10] F. Emannuel, "Solving Equal Piles with the Grouping Genetic Algorithm," Proceedings of the Sixth Intenational Conference on Genetic Algorithms, 1995.

[11] W. A. Greene, "Genetic Algorithms For Partitioning Sets," International Journal on Artificial Intelligence Tools, vol. 10, no. 01n02, pp. 225 - 241, 2001.

[12] S. Trichni, F. Omary, B. Boulahiat, and M. Bougrine, ”A new approach of mutation’s operator applied to the ciphering system SEC,” 6th ICCIT: International Conference on Computer Sciences and Convergence Information Technology (ICCIT 2011), 2011.

[13] M. Bougrine, F. Omaiy, S. Trichni, and B. Boulahiat, ”New evolutionary tools for a new ciphering system SEC version,” 2012 IEEE International Carnahan Conference on Security Technology (ICCST), 2012.

[14] Z. Kaddouri, F. Omary, A. Abouchouar , and M. Daari, ”Balancing Process to the Ciphering System SEC,” Journal of Theoretical and Applied Information Technology, 2013.

[15] A. Ray and D. Das, ”Encryption Algorithm for Block Ciphers Based on Programmable Cellular Automata,” Communications in Computer and Information Science Information Processing and Management, pp. 269 - 275, 2010.

[16] J. Bhaumik and D. R. Chowdhury, ”Design and implementation of Cellular Automata based diffusion layer for SPN-type block cipher,” 2012 International Conference on Informatics, Electronics & Vision (ICIEV), 2012.

[17] S. Roy, S. Nandi, J. Dansana, and P. K. Pattnaik, ”Application of cellular automata in symmetric key cryptography,” 2014 International Conference on Communication and Signal Processing, 2014.

[18] R. K. Mehta and R. Rani, ”Pattern generation and symmetric key block ciphering using cellular automata,” 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016.

[19] S. Bouchkaren and S. Lazaar, ”A New Cryptographic Scheme Based on Cellular Automata,” Lecture Notes in Electrical Engineering Proceedings of the Mediterranean Conference on Information & Communication Technologies 2015, pp. 663 - 668, 2016.

[20] S. Bouchkaren and S. Lazaar, ”A fast cryptosystem using reversible cellular automata,” International Journal of Advanced Computer Science and Applications, vol. 5, no. 5, 2014.

[21] K. M. Faraoun, ”A genetic strategy to design cellular automata based block ciphers,” Expert Systems with Applications, vol. 41, no. 17, pp. 7958 - 7967, 2014.

[22] K. Li, M. Sun, L. Li, and J. Chen, ”Image Encryption Algorithms Based on Non-uniform Second-Order Reversible Cellular Automata with Balanced Rules,” Intelligent Computing Theories and Application Lecture Notes in Computer Science, pp. 445 - 455, 2017.

[23] A. Y. Niyat, M. H. Moattar, and M. N. Torshiz, ”Color image encryption based on hybrid hyper-chaotic system and cellular automata,” Optics and Lasers in Engineering, vol. 90, pp. 225 - 237, 2017.

[24] Y. Wang, Y. Zhao, Q. Zhou, and Z. Lin, ”Image encryption using partitioned cellular automata,” Neurocomputing, vol. 275, pp. 1318 - 1332, 2018.

[25] K. Chakraborty and D. R. Chowdhury, ”CSHR: Selection of Cryptographically Suitable Hybrid Cellular Automata Rule,” Lecture Notes in Computer Science Cellular Automata, pp. 591-600, 2012.

[26] L. Mariot, ”Cellular Automata, Boolean Functions and Combinatorial Designs,” dissertation, 2018.

[27] J. Daemen and V. Rijmen, ”The Advanced Encryption Standard Process,” Information Security and Cryptography The Design of Rijndael, pp. 1 - 8, 2002.

[28] Robert G. Brown’s General Tools Page. [Online]. Available: https://phy.duke.edu/ rgb/General/dieharder.php.

[29] C. E. Shannon, ”A Mathematical Theory of Communication,” Bell System Technical Journal, vol. 27, no. 4, pp. 623 - 656, 1948.

[30] A. Biryukov and C. Canniere, “Linear Cryptanalysis for Block Ciphers,” Encyclopedia of Cryptography and Security, pp. 351 - 354, 2011.

[31] E. Biham, ”Differential Cryptanalysis,” Encyclopedia of Cryptography and Security, pp. 147 - 152.

Performance Analysis of Machine Learning Techniques for Smart Agriculture: Comparison of Supervised Classification Approaches

Rhafal Mouhssine[1], Abdoun Otman[2], El khatir Haimoudi[3]
Computer Science Department, Laboratory of Advanced
Science and Technologies
Polydisciplinary faculty, University UAE, Larache, Morocoo

***Abstract*—Agriculture form one of the most important aspects of life necessities, it is responsible to feed 7.7 billion person for the time being, and it is expected to supply more than 9.6 billion individual in 2050, the thing that made classical farming insufficient, and give birth to the notion of smart farming, and the race has begun toward using the latest technologies in the field. They integrate the Internet of Things (IoT), automation, Artificial Intelligence (AI), etc. And as researchers from a country that highly depends on agriculture, we have decided to also contribute to this evolution, and we chose Machine learning (ML) as our entrance to the field to satisfy the need for automated classification of the different products produced by a farm. In this work, we wanted to solve the problem of automatic classification of agricultural products, without the need of any human intervention, and we concentrate on the classification of red fruits, due to our proximity to a location that its product is red fruits. In other words, we are doing a comparative study among the well-known approaches that are used in image classification, and we are applying the best-found method to correctly classify the pictures of red fruits. And this empirically leads us to achieve great results as shown in the numerical result area.**

Keywords*—Support vector machine; K-nearest neighbor; deep neural networks; convolutional neural networks; smart agriculture; Cifar10

I. Introduction

The agriculture plays an important role in the economic systems of several countries, and one of these is our country, the Kingdom of Morocco, the agriculture forms one of the most important incomes to the country. Thus, increasing the effectiveness of the farming would also affect positively the economy of the kingdom, and develop somethings means to integrate the latest existing technologies in the field. And After the last revolution of the AI appears a term called smart farming, which directly affect the field of agriculture. But this short term assembles many intelligent technologies, and some of them already in use. But in this work, we chose to enter this world by using computer vision and image classification and use it to automatically classify the different species by the means of images.

Image classification is the ability to choose a unique correct label to the input image from a predefined set of categories, and it's considered one of the core problems in computer vision which resides in the intersection of several fields of studies: Mathematics, image processing, data mining, etc. Image classification has a large variety of applications such as object detection, segmentation, facial recognition, etc. and those applications can be used in larger practical applications like Surveillance Autonomous vehicles. Its complexity and its effectiveness highly depend on the method used to solve the problem since there are several methods that can be used to solve that problem (image classification).

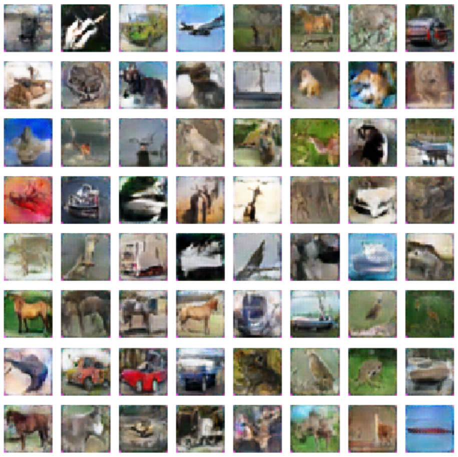


Fig. 1. Cifar10 dataset.

There have been several attempts to automate the process of image classification, but have chosen the closest papers to our work. In this area, Y. Abouelnaga and al tried on their work to work on an assembled model that use several CNN models and combine it with a KNN approach optimized by PCA (principal component analysis), and they have achieved good results on classifying the CIFAR-10 (Fig. 1) dataset [2]. In the same area, L. H. Thai et al. have used SVM together with artificial Neural networks to construct their model. They use feature-based sub-images and feed them to neural networks, and they use the SVM as the last layer that receives the results of the neural nets. And this approach made them reach a precision of 86% by applying their model on classifying human numerals [3]. As one of the first attempts at using convolutional neural networks Y. leCun et al. are one of the first ones who use convolutional layers and subsampling in order to extract the right features from images, even if the shape of the object inside the image has a large range of variance. Such as handwriting, and this made them achieve great results and inspire all the later CNN users on both image classification and NLP [4]. And to optimize the speed of training a neural Net and its variants Sergey Ioffe et al. make the normalization of each layer, the thing that made each layer learn independently. This addition reduces the overfitting, enabled the use of a higher learning rate and consequently makes the training faster and also enables the use of a larger number of layers [13]. Also in the agricultural area, Horea Muresan and Mihai Oltean have collected a new high-quality dataset, concentrated only on fruits named fruit-360, and to prove the quality of their dataset they apply a CNN based classifier, and they have got great results [19].

In this work, our objective will be to make a comparative study between the well know methods that attempted to solve the problem of image classification and to be more specific we will use K-nearest neighbor, Support vector machine classifier, Deep neural networks, and Convolutional neural networks, and after each implementation, we will mention the strengths and weaknesses of each method. and it's worth mentioning that all our tests will depend on the well known Cifar10 dataset, since their images have small dimensions (32X32X3), and it will let us experiment with our tests without the need for the clouds and expensive hardware. Despite the fact that it's hard to achieve good results with such highly pixelated images. And after choosing the right classifier and prove it by results we will apply it on our main problem which is the classification of red fruits, seeing their our importance to our country and especially to our country.

The rest of this paper will be organized as follows: we will begin by a study case section and, in the coming section, we will describe the K-Nearest neighbor its formal implementations, its applications on image classification and the results achieved with it as well as its weaknesses. Then we will devote the next section to the shallow learning method (Support vector machine) and its performance on the cifar10 dataset. After that, we will study the uniform neural networks, the difficulties to build a robust deep neural network and its performance in the same dataset. And we will leave the last section to the strongest method which is the Convolutional neural network and its performance of cifar10, and finally, we will conclude by a global conclusion which summarizes our work and gives an idea about our future perspectives.

II. Study Case

The kingdom of morocco depends largely on agriculture and it's one of the principal incomes of the country, according to Wikipedia, the agricultural sector in morocco accounts for approximately 13-15% of GDP (gross domestic product) as shown in (Fig. 2) and employs about 40% of the national workforce, and if we take the year 2011 as an example, we find that Morocco's GDP is 221 billion dollars and the agriculture has contributed to it by 15% [1].

Thus, improving the Quality or Quantity of agriculture directly affect the GDP of the country. For this reason, farming and agriculture, in general, is a strong power that can effectively ameliorate the income of the country. And for this work, we are trying to enter this interesting sector by the gate of smart farming, and we are trying to take advantage of our proximity to an agricultural area that takes a special care of the red fruits, and the possibility that we can enough information about the subject, to orient our objective to classify the different species of the red fruits after studying and analyzing the different existing approaches, and dedicate the best-found method to our case study.

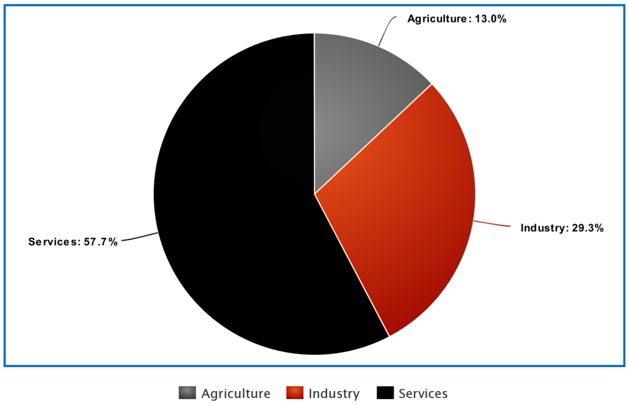


Fig. 2. GDP of Morocco.

III. K-Nearest Neighbor Classifier

K-nearest neighbor is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions, Fig. 3). A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common among its K nearest neighbors measured by a distance function.

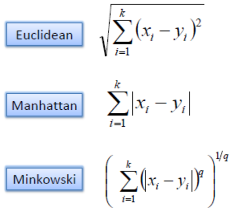

Euclidean
$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$
Manhattan
$\sum_{i=1}^{k}|x_i - y_i|$
Minkowski
$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

Fig. 3. Distance functions.

if K = 1, then the case is simply assigned to the class of its nearest neighbor. In general, a large K value is more precise as it reduces the overall noise. To choose the optimal value for K is best done by first inspecting the data, and cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value, but historically, the optimal k for most datasets has been between 3-10. that produces much better results than 1NN.

A. K-Nearest Neighbor and Image Classification

The K-Nearest neighbor classifier is by far the most simple machine learning classifier used in image classification. This machine-learning algorithm doesn't actually learn anything. To use it we simply flatten the images and turn them into vectors before passing them to the kNN classifier, which simply keeps them in memory without any processing. In other words, it keeps all the hard work to the prediction step. When we ask the classifier to predict the class of a new image, it calculates the distance between that new image and all the datasets which already kept in memory using one of the distance functions. Then it chooses the most k similar ones to the image. Next, it decides which class is more suited for this image.

B. Experiments and Results

In the beginning, it's hard to decide the most suited value of k to our problem nor the most effective distance function. So in our experiments, we used one of the most well-known methods which is cross-validation to decide the most effective hyperparameters to our problem (image classification), but we didn't focus that much on the distance function we just used the most known one (euclidean distance). After running several tests and experimenting several values of k we found approximately the same results for k between 3-10, but the best accuracy we have got is approximately 29%. But according to the best-known results, K-Nearest neighbor can reach 35% accuracy if it's used with the right distance function and right value of K. and there are also other ways such as principal component analysis which could improve its performance furthermore. Moreover, KNN nearest neighbor could be used in combination with convolutional neural networks to increase its accuracy[2].

C. Limitations

This approach has several flaws. Apart from its low accuracy, it also suffers from the extensive memory usage, which means that with a large dataset we will have problems to store the dataset, and also it has another major flaw, it has to do all the work in the prediction time so that the user must wait for the classifier to compare its image to all the dataset and calculate the distance between them and give it the most k nearest classes to the image, and this kind of behavior is not acceptable in real-time applications.

IV. Parameterized Classifiers

Using parameterized classifiers (Fig. 4) helps us overcome the major flaw of K-Nearest Neighbor because in this case all the time-consuming tasks are done in the training stage. Once the training is complete, we can discard all the training dataset and free the memory, we just preserve the learned parameters W and b. And since we can have these parameters (w and b), we quickly predict the new test data since all we have to do is a simple linear transformation:

$$f(x_i, W, b) = Wx + b$$

A. Train a Linear Classifier

To train this type of models we only need to adjust the parameters W and b in a way that helps us achieve the best possible accuracy, and we do accomplish so with help of a loss function (quantifies how well our prediction agree with the ground-truth label) which we try to minimize using Stochastic gradient descent or one of if its variants.

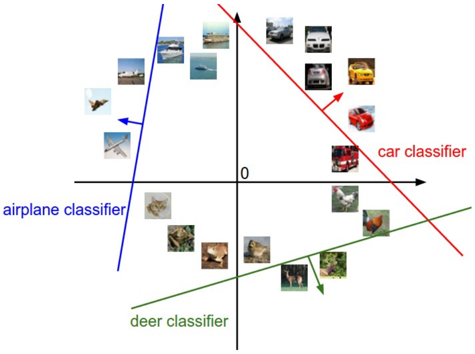


Fig. 4. Linear classfier.

B. Loss Function

The loss function is one of the most important pieces of all parameter based classifiers, and as we have previously mentioned, the loss function tells us how good our prediction compared to the ground truth label [6]. For linear classifiers, we can use multiple loss functions, but the most commonly used are this two:

- Multi-class Support vector machine, also known as hinge loss: inspired from the famous support vector machine classifier [5]:

$$s_j = Wx_i + b$$
$$L_i = \sum_{j \neq y_i} max(0, s_j + s_{y_i} + 1)$$

- cross-entropy: which used with softmax (Fig. 5) classifier that uses probabilities to describe the confidence of each class:

$$s_j = Wx_i + b$$
$$L_i = -\log(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}})$$

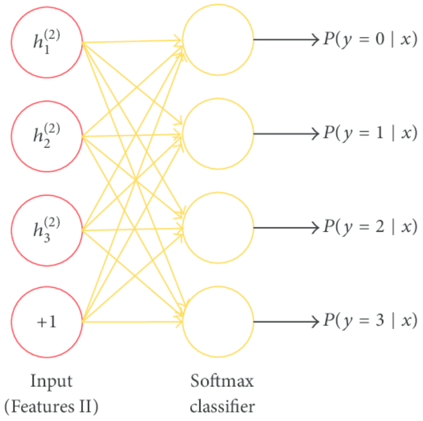


Fig. 5. Softmax classifier.

After calculating the loss function L_i of each example in the batch (or minibatch), we do calculate its mean $L = \frac{1}{n}\sum_{i=1}^{n} L_i$, to get a more global view, and converge in the direction of all the training examples, instead of zigzagging in the direction of each training example at each iteration and consequently slow down the training.

C. Limitations

Although linear classifiers are much better than K-nearest neighbor method and overcomes most of its flaws, in terms of accuracy they still suffer, in our case we have tried the two versions of linear classifiers (the svm based, and softmax classifier), we didn't exceed 45% in both cases, despite the all the efforts we did to tune parameters and experiment with the combinations of hyperparameters.

V. Neural Networks

Neural networks are the most effective machine learning algorithm, and it can easily outperform almost any other machine learning algorithm in any task that involves learning, and its architectures has a wide range of variants (DNNs, CNNs, RNNs, AutoEncoders, GANs…), which make it makes it capable to perform a large variety of tasks such as Object detection, Image recognition, Regression, compression …, and it's used in almost any modern applications that require some sort of intelligence. Even the most simple form of a Neural network (shallow Network) which consists of only two layers (hidden, output), is considered as a universal function, and in theory, it could approximate any existing mathematical function.

Neural networks share a lot of the common notions of the classical methods of machine learning (especially the ones that uses trainable parameters) such as normalization, loss functions, activation functions, optimization techniques (gradient descent, stochastic gradient descent), but Neural Nets are characterized by another type of notions that are specific to them like the fact that they could contain a large number of layers, and that they use backpropagation to train an arbitrary number of layers, which make them special and give them high flexibility that enables them to adjust to any kind of data. The powerful architectures of Neural Nets made them prove their effectiveness and attract the curiosity of the researchers which consequently made them one of the most active research areas. They have focused on every detail of Neural Nets. There are researches in weight initialization, activation functions, regularization, normalization, and even in the right number of layers. Thus, in this work, we have included the most recent terms and tried to use the latest studies and the best choices to construct our own neural network and use it to classify the Cifar10 dataset, and the following subsections describe the elements that we have used in our implementations.

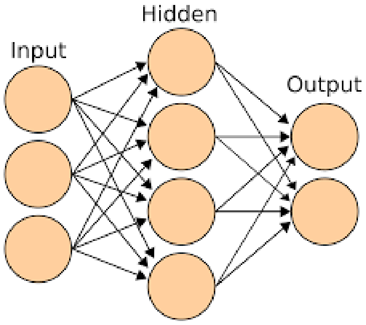


Fig. 6. Simple shalow network.

A. Regularization

Neural networks are considered the most flexible machine learning algorithms and can adapt with any type of data as discussed previously, but this flexibility comes with a cost: Overfitting (Fig. 7), In other words, they memorize the training data which make them unable to generalize and recognize new data, and that's where the term of regularization could help to prevent this Phenomenon.

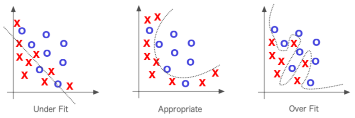


Fig. 7. Overfitting phenomenon.

There two major types of regularization L2 regularization and Dropout, there's also L1 regularization but its not preferable.

1) L2 regularization: is the most known, and it's not exclusive to neural networks, it's also used with a large variety of machine learning algorithms, it's simply an addition of the Frobenius norm of the weight matrix to the loss functions, which decays the weights and consequently encourages the simple version of the neural network model, the thing that prevents overfitting to some extent. and since the L2 regularization encourage small weights, it also does another important job that serves positively some non-linearities such as sigmoid and tanh, because it confines the weights in the small portions where it can make use of the linear area of the non-linearities as shown in (Fig. 8), this thing accelerates the learning process because the gradient isn't dead, in contrast to areas where $|w|$ are large.

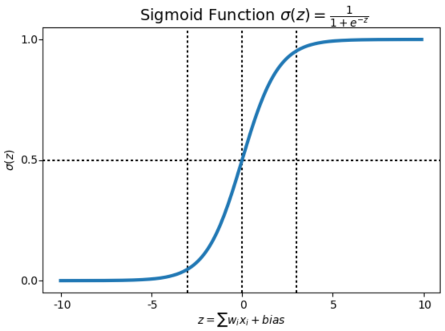
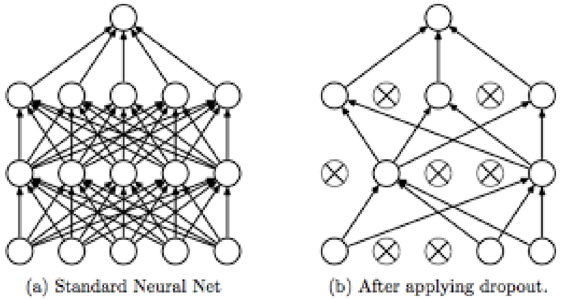


Fig. 8. Linear portion of the sigmoid

2) Dropout: does also a similar job and make each time a simple version of the neural network, by deactivating a number of neurons on each layer according to so some pre-specified probability (Fig. 9) and it does the work because it prevents the model to rely on any specific feature, and make it take different paths each time so that it can finally generalize well [10].


(a) Standard Neural Net

(b) After applying dropout.

Fig. 9. Dropout

B. Input Normalization

Input normalization/standardization is a simple preprocess of the input data, that can be summarized by the following equations:

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{1}$$

$$\sigma^2 = x_i \tag{2}$$

$$\hat{x_i} = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{3}$$

Input normalization is an important process for a large variety of machine learning algorithms. It does work because it prevents the large variance between input features, which could cause the data to be sensible to the parameter update and risk to make the gradient overshoot in some directions. Input normalization also makes the input data zero mean, which inhibits the parameter update to be in the same direction.

C. Optimization

The optimization is the most important building stone of the learning process of a neural network. Almost always meant by optimization gradient descent and its variants in the context of the literature of neural network training, we could train our model using another kind of optimizers such as nature-inspired algorithms(meta-heuristics), but by far the gradient descent and its variants are the most suited to train neural nets. The original version of gradient descent is considered to be too slow because at each iteration it needs to explore all the training data to make one step. As a major successor of gradient descent is the mini-batch stochastic version of gradient descent, which doesn't need to traverse all the data, it just takes a random prefixed amount of training data from the training set and evaluate the loss and then take a step. Although, SGD make a good job in replacing Gradient descent, it has some drawbacks. SGD makes a parameter update with just a subset of the training set, which makes the direction of the update has some variances, and thus, the path taken by SGD, will oscillate toward convergence, and those oscillations forces us to use a small learning rate which consequently slows down the learning process. For this reason, there were several attempts to solve this problem, the most famous ones are: sgd+momentum [7], RMSprop [8], and Adam [9].

D. SGD + Momentum

The idea behind SGD+momentum is that it adds a little momentum to the gradient, by adding a new term called velocity, which is simply the exponentially weighted average of the gradient. In one side it helps us escape from the critical points where the gradient could die, and in the other side it tends to average out the oscillations in the directions that aren't towards minima, which make by making them smooth, and since the motion, toward the minima, is stable it doesn't affect the velocity toward convergence, instead, it accelerates the learning process and it allows us to use larger learning rate. The following equations are the simple modification made to the update when we use SGD+momentum:

$$V_{\partial W} = \beta V_{\partial w} + (1-\beta)\partial W \tag{4}$$

$$V_{\partial b} = \beta V_{\partial b} + (1-\beta)\partial b \tag{5}$$

$$W = W - \alpha V_{\partial W} \tag{6}$$

$$b = b - \alpha V_{\partial b} \tag{7}$$

E. RMSprop

RMSprop is from the family of adagrad (adaptive gradient) optimizers. This type of optimizers tries to adjust the learning rate of each parameter independently, by performing smaller updates to the frequently occurring features, and larger updates for parameters associated with infrequent features, the thing that make them able to handle sparse data very well, but given the cumulative nature of the term that tries to adapt the learning rate, it creates the problem of continuously decreasing the learning rate, which leads to halting the learning process. That's why RMSprop along with other algorithms come as extensions to the original adagrad algorithm, as an attempt to fix this disadvantage, in the case of the RMSprop, it simply tries to replace the accumulative term by a running average which makes it decay with time, and forget about the old values as shown in following equations:

$$S_{\partial W} = \beta S_{\partial W} + (1-\beta)\partial W^2 \tag{8}$$

$$S_{\partial b} = \beta S_{\partial b} + (1-\beta)\partial b^2 \tag{9}$$

$$W = W - \alpha \frac{\partial W}{\sqrt{S_{\partial W} + \epsilon}} \tag{10}$$

$$b = b - \alpha \frac{\partial b}{\sqrt{S_{\partial b} + \epsilon}} \tag{11}$$

F. Adam

Adaptive momentum is one of the most effective algorithms that used to optimize NNs, and recently it becomes the standard. Adam optimizer doesn't reinvent the wheel, instead, it's simply a combination of the concepts of the two previously discussed optimizers, it takes advantage of both of them, and it almost always outperforms them in practice. and the following equations show how it combines the two set of equations of the SGD+momentum and RMSprop:

$$V_{\partial W} = \beta_1 V_{\partial w} + (1-\beta_1)\partial W \tag{12}$$

$$S_{\partial W} = \beta_2 S_{\partial W} + (1-\beta_2)\partial W^2 \tag{13}$$

$$\hat{V}_{\partial W} = \frac{V_{\partial W}}{(1-\beta_1^t)} \tag{14}$$

$$\hat{S}_{\partial W} = \frac{S_{\partial W}}{(1-\beta_2^t)} \tag{15}$$

$$W = W - \alpha \frac{\hat{S}_{\partial W}}{\sqrt{\hat{S}_{\partial W} + \epsilon}} \tag{16}$$

There are more optimizer and more alternatives, that we haven't discussed here such as nestrove algorithm which is an extension to sgd+momentum, and we didn't mention them because they aren't used in practice, but as shown in (Fig. 10), the Adam optimizer is the most powerful.

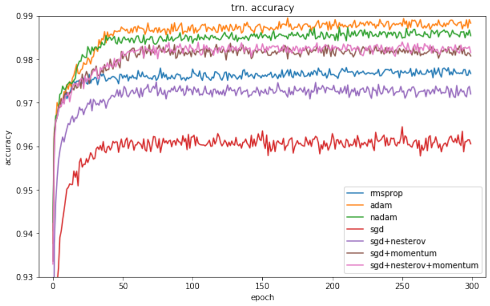


Fig. 10. Optimizers comparison

G. Deep Neural Networks

Deep neural networks are simply a version of neural networks (Fig. 6) with more than one hidden layer (Fig. 11). In principle, you don't need a deep neural network. And given enough training data, a large neural net with only one hidden layer can approximate any mathematical function. But the problem with extremely large single hidden layered neural networks is the lack of generalization, they could memorize but this is not enough. If we test a super-wide shallow network with new data, it won't do well, even if it could memorize all the training data. and this is not useful in a real-world scenario.

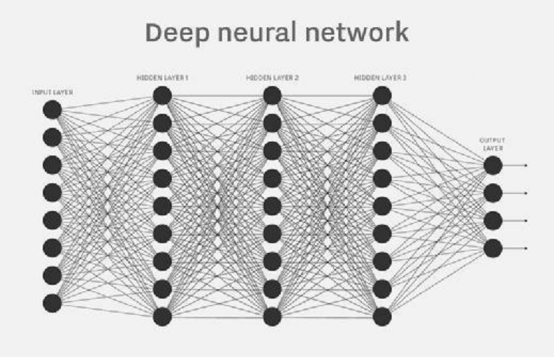


Fig. 11. Deep Neural Network

In the other hand, a deep neural net with multiple hidden layers learns in a different way, the first layers learn to recognize basic things such as edges in the case of pictures, and the deeper layers learn more complicated things that are constructed from combinations of the things learned in the earlier layers, and this gives multi-layered neural nets the ability to generalize better, and this serves better a practical application. Deep neural networks are extremely useful, they generalize well, learn better and achieve better results, but the complexity in there architecture comes with a cost, they are hard to train in comparison to the other simple shallow networks, deep neural networks use the same principle as the other regular networks, but in case of DNN, there are some other things that should be considered, like the weight initialization and batch normalization to simplify the learning process.

H. Weight Initialization

One of the starting points to take care of while building your network is to initialize your weight matrix correctly. Weight initialization also plays an important role in training Deep Neural networks, it might seem evident, and we might think that we could initialize weights with just some random values or just initialize them with zero. But it's not that simple, if we do initialize them with zero, we will get the same output results, and eventually get the same results which will lead us to update the weights with the same values, and also if we think to initialize them with the extremely small values we will risk having weights decays in deeper layers as shown in (Fig. 12), and if we initialize them with large numbers, we will suffer from having vanishing gradients especially with some non-linearities such as sigmoid and tanh.

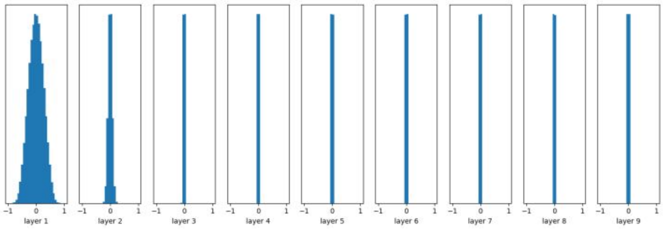


Fig. 12. Weight decay in deep layers

That's why weight initialization is considered to be an important task and gets the attention of many researchers, and it's one of the widest areas of researches that concerns neural networks, but the most known two methods are the one called xavier initialization [11], and another extension [12] to it that works better with relu activation variants. And the initialization equation of those two methods are as follows:

$$W^{[l]} = random(W)(from - normal - distribution) * \frac{1}{\sqrt{n^{l-1}}}$$

$$W^{[l]} = random(W)(from - normal - distribution) * \frac{2}{\sqrt{n^{l-1}}}$$

where n^{l-1} is the size of the input from the previous layer.

I. Batch Normalization

Training Deep Neural Networks is complicated Due to the differences in distributions of the inputs of each layer caused by the constant changes of the parameters of previous layers, and this makes training process becomes too hard, and to make progress we have to lower the learning rate and be too careful when we initialize the parameters, the thing that make the training too slow. We refer to this phenomenon as an internal covariate shift and address the problem by normalizing each layer's inputs (Fig. 13). When we introduce normalization and normalize each training mini-batch, we can use a larger training rate and be less cautious about the initialization process. And it also acts as a regularizer. Also, batch normalization allows each layer of a network to learn by itself a little bit more independently of other layers.

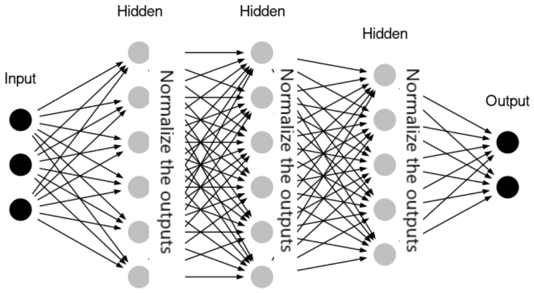


Fig. 13. Batch Normalization

Batch normalization normalizes the output of a previous activation layer by subtracting the batch mean and dividing by the batch standard deviation. However, after this shift/scale of activation outputs by some randomly initialized parameters, the weights in the next layer are no longer optimal. SGD (Stochastic gradient descent) undoes this normalization if it's a way for it to minimize the loss function. Consequently, batch normalization adds two trainable parameters to each layer (Fig. 14), so the normalized output is multiplied by a "standard deviation" parameter (gamma) and add a "mean" parameter (beta). In other words, batch normalization lets SGD and its variants do the denormalization by changing only these two weights for each activation, instead of losing the stability of the network by changing all the weights.

The Batch normalization operation is simply governed by the following equations [13]:

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β
Output: $\{y_i = \text{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \text{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Fig. 14. Batch Normalization equations

Where γ and β are trainable variables.

We have implemented this notion in our own version of the neural network, and experiment with different sizes of batches to see the differences that batch normalizes make and how it accelerates the training phase, and we have summarized our experiments in (Fig. 15).

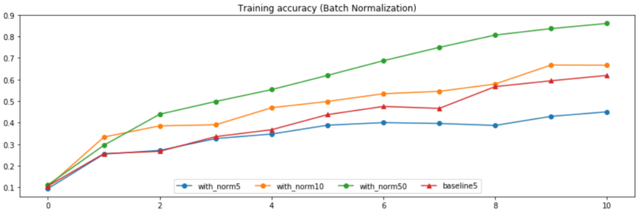


Fig. 15. Batch Normalization Effect

As shown in the figure above, despite the clear acceleration of the training, it's clear that the size of the batch has a huge effect on the effectiveness of batch normalization which could make a problem, that's why there where several attempts to solve it, such as layer normalization [14] and also recently appears an activation function that's his author claims that it eliminates the need of the batch normalization [15], but all those claims are still under test and the batch Norm still prove its effectiveness for the time being.

J. Deep Neural Networks and Cifar10

After implementing our own version of the neural network, that we have tried to insert all the above-discussed concepts in it, and we've tried to experiment with the best possible options, we have adjusted our network to classify Cifar10 dataset, that we chose it to be our criterion of the performance of our classifiers. Then after building the most convenient version and after a series of hyperparameters tuning to find the best combinations of hyperparameters, we test and we get an accuracy that exceeds 55%. And that's kind of disappointing after all this work, but that's happened because we've ignored the fact that the dataset is images. And here comes the role of another type of Neural Network called Convontional neural networks, which more suited to this kind of dataset (images).

VI. Convolutional Neural Network

Convolutional Neural network is just a deep Neural network with a different structure, and it has been proven empirically that CNN is by far the most effective Neural Network architectures for image classification, it even outperforms the human performance on classifying the imageNet dataset on 2015 [12]. This kind of powerful performance on image recognition tasks enabled the CNNs to achieve great results in bigger use cases such as object detection, and segmentation. The Convnets benefit from all the features and ideas of the usual deep neural network, they use all the techniques explained in the previous section. But they use two additional layers (convolution and pooling), which are the real cause behind the outstanding performance of CNNS.

A. CNN History

The concept that has lighted the idea of Convolutional Neural Networks has begun decades ago with the conclusions of the two famous research papers titled: “Receptive Fields of Single Neurons In The Cat’s striate cortex” in 1959 and “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex” in 1962, made by Hubel and Wiesel. they observed that the neurons have a hierarchical organization, and that earlier layers response to light orientations, and later layers response to light orientations and movements, and the last set of layers (that contains the most complex types of neurons) responds to movements and endpoints. Then in 1980, Fukushima [16] has built the first example of a network architecture model that has this idea of simple and complex cells. afterward, in 1998 they built the first model that applies backpropagation and gradient-based learning to train a CNN and they were able to do a good job on document recognition [17], and also did well on digit recognition, but it wasn’t able to scale to more complex data until the appearance of the AlexNet [18] in 2012, which has been able to achieve great results, scale to larger and complex data, and make use of the latest hardware.

B. Convolutional Layer

The convolutional layer is the most important part of a convolutional network and it’s the one that does contain the most valuable set of parameters. The Convolutional layer’s parameters consist of a set of learnable filters (also called kernels or feature detectors); every filter contains a small set of weights spread vertically and horizontally and through a specified depth (Fig. 16). A usual size of the first layer is 5X5X3 or 3X3X3.

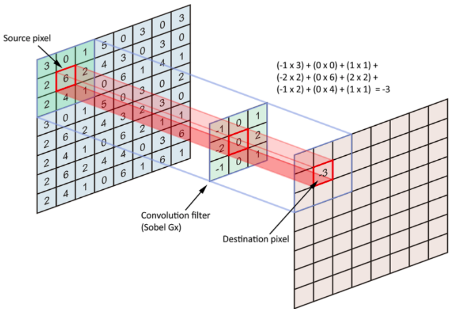
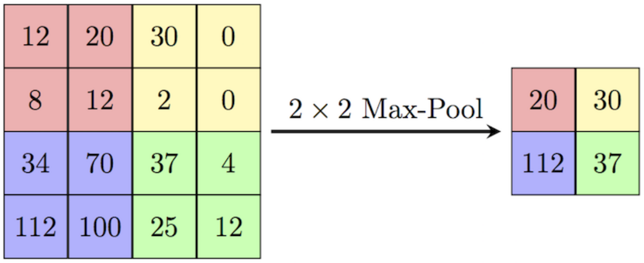

Source pixel
(-1 x 3) + (0 x 0) + (1 x 1) +
(-2 x 2) + (0 x 6) + (2 x 2) +
(-1 x 2) + (0 x 4) + (1 x 1) = -3
-3
Convolution filter
(Sobel Gx)
Destination pixel

Fig. 16. Convolution operation

To perform the convolution operation, we slide each filter across the width and height of the input volume and compute the dot products between the entries of the filter and the input at any position, and this operation will produce a 2-dimensional activation map that gives the response of that filter at every spatial position, and eventually, through training process the network will learn filters that activate when they see some type of visual feature.

C. Pooling Layer

The pooling layer reduces the number of parameters and calculations in the network. Thus, it improves the efficiency of the network and avoids over-learning. The pooling layer receives several feature maps and applies to each of them a pooling operation, which used to reduce the size of the images while preserving their important characteristics. For this, the image is cut into regular cells, then the maximum value is kept within each cell (Fig. 17). In practice, small square cells are often used to avoid losing too much information. The most common choices are adjacent cells of size 2X2 pixels that do not overlap. The same number of feature maps is preserved, but these feature maps are much smaller.



12	20	30	0
8	12	2	0
34	70	37	4
112	100	25	12

20	30
112	37

Fig. 17. Pooling operation

Thus, the pooling layer makes the network less sensitive to the position of features: the fact that a feature is a little higher or lower, or even that it has a slightly different orientation should not cause a radical change in the classification of the image.

D. CNN Architectures

CNN has proven that it is the most efficient image classifier since the 2012 imageNet international competition using AlexNet architecture [18], which is very similar to LeNet architecture which used in 1998 for digit recognition, But AlexNet after 14 years was able to take advantage of the computational power of GPUs, and consequently could be used for more powerful and realistic datasets, such as ImageNet. And after the great performance achieved at that time by AlexNet, the world’s attention has turned again toward ConvNets, and all the subsequent winners used one of the variants of CNN.

Until now, dozens of CNN architectures have appeared, But the top architectures that we’re able to positively affect the evolution of CNN were 3: VGGNet 2014, GoogLeNet 2014, ResNet 2015.

1) VGGNet (Visual Geometry Group Net): was ranked second in 2014 competition, but it used a distinctive interesting idea with regard to the receptive field of the filters used in convolutional layers, so instead of using 5X5, 7X7 or 11X11 like in the case of AlexNet, VGG uses only two 3X3 receptive fields to replace the 5X5 filter, and five 3X3 to replace the 11X11 receptive field (Fig. 18). This way it could effectively replace the large filters without hurting the performance, in the case of 11X11 filter we get 121 parameters, and VGG achieves the same results with only 3X3X5 = 45 parameters[20].

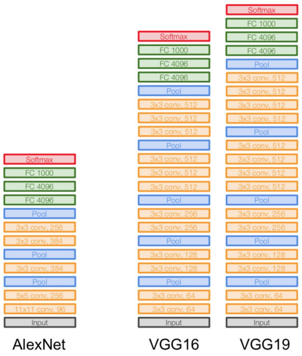


Fig. 18. AlexNet vs VGG

2) GoogLeNet: also known by the name of **Inception V1**, and it is the winner of 2014 competition, GoogLeNet is formed from a number of inception modules (Fig. 19), and each one of this modules contains a number of parallel convolutional layers, that uses filters with different receptive fields and a pooling layer in addition to a concatenation layer which sums up the output of the parallel layers depth-wise [21].

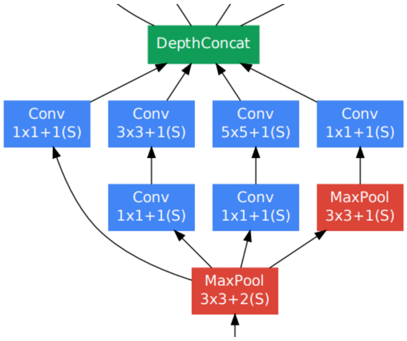


Fig. 19. Inception module

This way googleNet was able to increase the number of layers to 22, with 12 times less number of parameters in comparison to AlexNet.

3) ResNet: winner of 2015 competition, and the first one who outperform the human capability of classifying imageNet dataset (Fig. 20), with only 3.57 error.

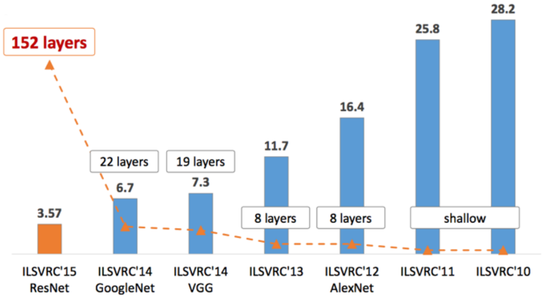


Fig. 20. Evolution of CNNs.

ResNet was able to dramatically increase the depth of the neural networks with an innovative idea which simplifies the $f(x)$ that needed to be learned by each layer (Fig. 21), this happens by adding an identity function to the residual $f(x)$, which means that the layer only needs to learn a $\Delta = H(x) - x$ [12].

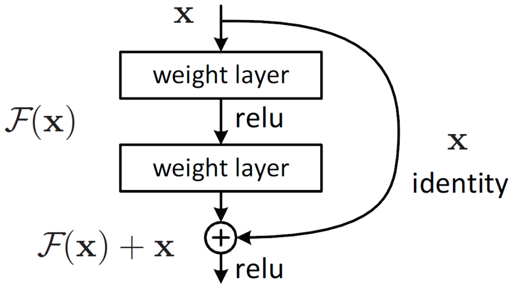

x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Fig. 21. Residual.

This way resNet was able to go very deep and use 152 layers. ResNet was also able to affect the normal deep neural networks and made them able to attain 1000 layers.

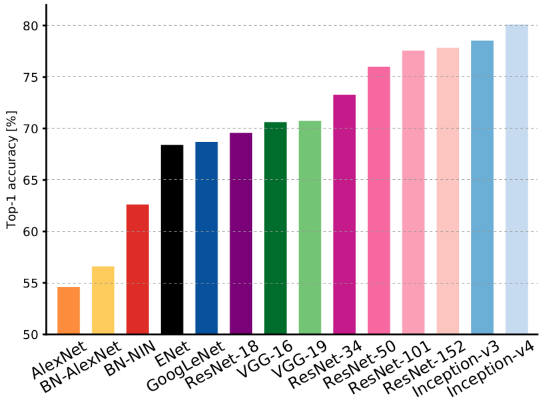


Fig. 22. Comparison of accuracy [22].

In the subsequent years, the architectures that could win the famous ILSVRC competition were only some kind of a hybridisation or reformulation of this architecture, and Inception v4 is an example of a hybridisation of resNet and googLeNet that gives the best performance in term of accuracy. And (Fig. 22 and 23) are an overview of the performance of this architectures.

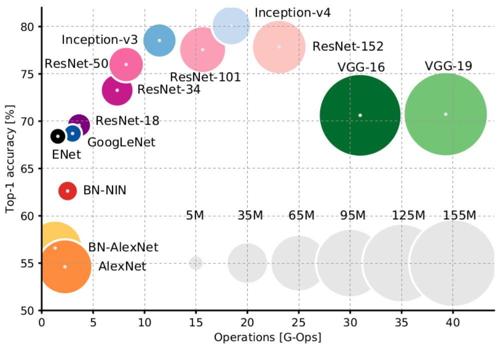


Fig. 23. General comparison of CNN architectures [22].

E. Experimental Results

After studying the theory behind convolutional neural networks, we have made an application that took advantage of all the notions discussed in the usual neural networks, and additional features of the convolution neural network, and after spending a fair amount of time searching we have found the hyperparameters and structure that would best suit our test dataset (cifar10) and give as a precision that exceeds 90% as shown in Table I.

TABLE I. CLASSFIER PERFORMANCES

Method	KNN	Linear Classifier	Neural Network	CNN
accuracy	28%	45%	55%	>90%

Despite our knowledge about the important types of architectures, and the way to achieve good results in image classification, we weren't able to experiment with these architectures and use them with Bigger and more realistic datasets such as imageNet, because of the lack of the adequate hardware. and we were forced to use the simple form of the convolutional neural network, but it was sufficient to get great results in the Cifar10 dataSet, and we also thought to apply these notions on a real use case that we could benefit from. So we chose to use it in agriculture and classify a dataset of red fruits since our nearby area (Larache city) is suitable for planting red fruits. Thus, we chose a subset (only red fruits Fig. 24) of a famous dataset that classifies fruits [19].

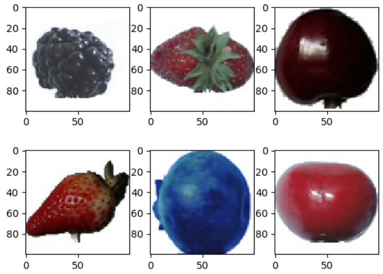


Fig. 24. Sample from red fruits set.

To accomplish this task, we adapt our convolutional neural net model to classify this subset by adjusting hyperparameters and preprocessing the raw images of the dataset. By doing so we have achieved a precision that reaches 99.9% because of the simple nature of the data set. And as presented in (Fig. 25) all the guesses of the model are correct.

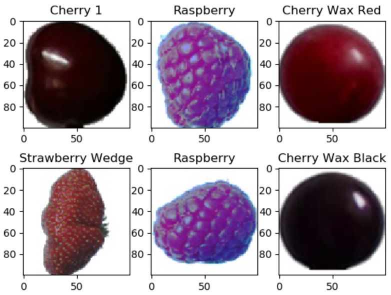


Fig. 25. Cnn model predections.

VII. CONCLUSION

In this work, we have experimented and tested a fair amount of classifiers that are used to recognize images, and we have known the strengths and weaknesses of each one of them, we have also studied in depth neural networks and convolutional neural networks and achieved good results in classifying our chosen datasets. And we are looking forward to do more, we want to explore all the variations of neural nets, and also apply them in more interesting fields of study like object detection and segmentation, and also use them in a real applications that could directly affect our everyday life.

REFERENCES

[1] Fanack, *Independent online media organization committed to publishing and disseminating balanced and informed analysis about the Middle East and North Africa*, consulted on 22-19-2019.

[2] Y. Abouelnaga, S. Ali, H. Rady, and M. Moustafa. *CIFAR-10: KNN-based Ensemble of Classifiers*. 2016.

[3] L. H. Thai, T. S. Hai, N. T. Thuy, *Image Classification using Support Vector Machine and Artificial Neural Network*, 2012.

[4] Y. LeCun, P. Haffner, L. Bottou, and O. Bengio, *Object Recognition with Gradient-Based Learning*, 1998.

[5] C. Cortes and V. Vapnik. *Support-Vector Networks*. In: Mach. Learn. pages 273-297. Sept. 1995.

[6] A. rosebrock. *Deep learning for computer vision with python*. 1st Edition., 2017. [online]. Available: https://www.pyimagesearch.com/deep-learning-computer-vision- python-book.

[7] N. Qian, *On the Momentum Term in Gradient Descent Learning Algorithms.*, In: Neural Netw. 12.1 , pages 145-151, Jan. 1999.

[8] S. Ruder. *An overview of gradient descent optimization algorithms*. 2017.

[9] P. Kingma, J. L. Ba, *ADAM: A method for stochastic optimization*. 2017.

[10] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. 2014.

[11] X. Glorot, Y. Bengio, *Understanding the difficulty of training deep feedforward neural networks*. 2010.

[12] K. He et al., *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*. 2015.

[13] S. Ioffe, C. Szegedy, *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015.

[14] J. L. Ba, J. R. Kiros, Geoffrey E. Hinton, *Layer Normalization*. 2016.

[15] G. Klambauer, T. Unterthiner, A. Mayr, *Self-Normalizing Neural Networks*. 2017

[16] K. Fukushima, *Neocognitron, A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position*. 1980.

[17] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, *Gradient Based Learning Applied to Document Recognition*. 1998.

[18] A. Krizhevsky, I. Sutskever and G. E. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*. 2012.

[19] H. Mureşan, M. Oltean, *Fruit recognition from images using deep learning*. 2018.

[20] K. Simonyan, A. Zisserman, *Very deep convolutional networks for large-scale image recognition*. 2015.

[21] C. Szegedy et al., *Going deeper with convolutions*. 2014.

[22] A. Canziani, E. Culurciello, A. Paszke. *An analysis of deep neural network models for practical applications*. 2017.

Minimal Order Linear Functional Observers Design for Multicell Series Converter

Mariem Jday[1]
Laboratoire de Recherche
en Automatique,
Ecole Nationale d'Ingénieurs de Tunis,
Université de Tunis El Manar
Le Belvédère, 1002 Tunis, Tunisia

Paul-Etienne Vidal[2]
Laboratoire de Génie
de Production,
Ecole Nationale d'Ingénieurs de Tarbes,
47 avenue d'Azereix,
65016 Tarbes CEDEX, France

Joseph Haggège[3]
Laboratoire de Recherche
en Automatique,
Ecole Nationale d'Ingénieurs de Tunis,
Université de Tunis El Manar
Le Belvédère, 1002 Tunis, Tunisia

***Abstract*—The requirement of high voltage-power level for many applications like the energy conversion system give rise to use a structure of a multilevel converter like the multicell series converter. To benefit as much as possible from this power converter an appropriate voltage distribution for each cell must be performed, hence the need to estimate this voltages. This paper aims to design a minimal single linear functional observer for a multicell series converter to estimate the capacitor voltages. Based on its hybrid model, an observability study prove the ability to estimate this capacitor voltages. Also a linear functional observers are proposed using a direct procedure without solving the Sylvester equation and based on an operation mode classification approach. Simulations of four-cells multicell converter are given in order to check the efficiency of the converter's hybrid model and the performance of the proposed minimal single linear functional observers.**

Keywords*—*Multicell converter; voltage capacitor; hybrid model; Z(TN)-Observability; functional observer

I. Introduction

Due to the energy transition combined with e-mobility, energy conversion systems require higher power level. This can be achieved with an increase of the functioning current or voltage. The increase of the functioning voltage is done in usual power structures with an increase of the switches's voltages. For instance, [1] illustrates how to apply 10 kV Silicon switches (Si) in a two voltage level converter. However, high blocking voltage switches have a lower dynamic performance. To mitigate both power level associated with higher blocking voltage, and good switching dynamics, Multicell Series Converter (MSC) are an alternative to usual conversion structures. In such a structure, the output leg voltage is produced by a specific association of switching cells. The main advantage of MSC is to increase the degree of freedom of the conversion structure and then to contribute to the reduction of the voltage constraints on the switches, such as the reduce of the harmonic distortion rate which lead to obtain a high quality output voltage [2], [3]. Nevertheless, MSC structures have inner capacitors. Each capacitor voltages must have a specific value according to the applied input voltage. This is done according to a suitable control strategy applied. A risk known of such a structure is the capacitor voltage imbalance [3], which should be avoided. As a matter of facts, capacitor voltage must be known at each time. This is possible by the mean of sensorless strategy, also called observers, instead of the extra sensors which can increase complexities and costs of converters. Due to the converter structure complexity, one drawback of sensorless techniques is the observability statement [4]. Hence, it is needed to proceed to an observability study before the observer design. The aim is to prove the possibility to estimate the voltage capacitors. For MSC, the observer design can be made with two types of model: the average model and the instantaneous model. In the MSC average model, the non linear behavior of the system is highlighted, so a non linear observer must be used such as the non linear sliding observer presented in [5].

The second model is the instantaneous hybrid model which expresses the dynamic of continuous variable according to the state of discrete variables.

Several types of observers based on this model are developed, such as the adaptive linear observer given in [6], [7] and the sliding mode observer presented in [8].

In order to reduce the estimation algorithm complexity, linear observers can be an interesting alternative. In order to avoid non-linearities observers should be designed from the instantaneous model. Among this observers, the linear functional observer (LFO) which is a low order observer is can be used, it allows to estimate a function of the state vector. In particular each capacitor voltage of the MSC can be estimated thanks a state function.

Recently, [9] detailed the design of the functional observer for switched discrete system.

Furthermore, it is noticed that the study of the functional observer for a multilevel converter (even for a continue switch system) has not been reported despite its utility for such system which present an observability problem caused by a switch behaviour. This paper tackle this problem. An observability study and an appropriate strategy to estimate the MSC voltages are given for an hybrid model of the MSC. Thereafter an observation strategy using the functional observer is defined allowing to isolate each capacitor voltage and reduce the complexity of the observer design strategy. The procedure to get the minimal order functional observer presented in this paper is based on the work of F. Rotella et al. [10], in which, it is demonstrated that the solution of Sylvester equation is not given by using this direct method. This work provides the ability to set an arbitrary dynamics to this observer to increase its performance.

The paper is arranged as follows. In Section 2 the hybrid model of the multi-cell series converter is given. Section 3 discusses the problem of the observability of the voltage capacitors. Section 4 is dedicated to the design of the functional observers for this system. Before the conclusion an application to a four-cells converter is provided including simulation results.

II. Model of Multicell Series Inverter

The multicell serie converter [11] [12] [13] was introduced in the 90's . The inverter structure of a MSC is illustrated in Fig. 1. The MSC is a serial connection of p pairs of switches $\{S_j, S_j^{'}\}$. The switches S_j and $S_j^{'}$ are grouped into switching cells where each switch S_j is associated to its binary state such as $S_j \in \{0, 1\}$, $S_j = \bar{S}_j^{'}$. $j \in \{1, \dots, p\}$, $j \in \{\mathbb{N}\}$.
I_{Ch} and V_{Ch} are respectively the load current and voltage , E is the constant input voltage, L and R are respectively the inductance and resistance of the load and V_s is the leg voltage. The inner voltage floating capacitors are $V_{C1}, ..., V_{Cp-1}$. As there is $p-1$ capacitors inserted, it exists $p-1$ capacitor voltages and currents such as:

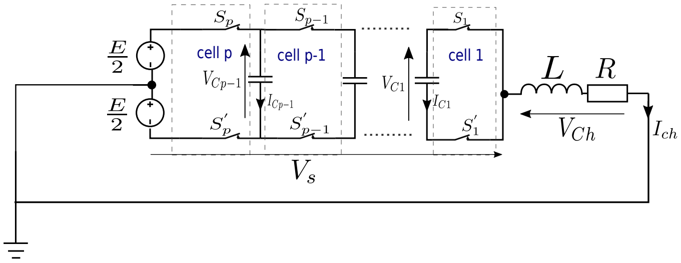


Fig. 1. Multi-cell series inverter.

$$V_{Cj} = \frac{jE}{p}, \tag{1}$$

$$I_{Cj} = I_{Ch} \times (S_{(j+1)} - S_j), \tag{2}$$

where j ponderates the input voltage E.

The relationship between the leg voltage V_s and the load voltage V_{Ch} is:

$$V_{Ch} = V_s - \frac{E}{2} = L\frac{dI_{Ch}}{dt} + RI_{Ch}. \tag{3}$$

Then, the load current I_{Ch} is:

$$\dot{I}_{Ch} = \frac{V_s}{L} - \frac{R}{L}I_{Ch} - \frac{E}{2L}. \tag{4}$$

As stated in [**?**], equations (3) and (4) allow to express the leg voltage V_s as:

$$V_s = ES_p + \sum_{j=1}^{p-1} V_{Cj}(S_j - S_{(j+1)}). \tag{5}$$

Indeed, the new differential equation of the load current is defined by:

$$\begin{aligned} \dot{I}_{Ch} = &-\frac{R}{L}I_{Ch} + \frac{S_pE}{L} - \frac{E}{2L} \\ &-\sum_{j=1}^{p-1}\frac{V_{Cj}}{L}(S_{(j+1)} - S_j). \end{aligned} \tag{6}$$

Subsequently, the dynamic behavior of the voltage capacitor V_{Cj} is expressed as follow:

$$\dot{V}_{Cj} = I_{Ch}\frac{1}{C_j}(S_{(j+1)} - S_j). \tag{7}$$

$d_j = S_{(j+1)} - S_j$ and $d_p = S_p$ will be further introduced in order to simplify the expression. Let us remark that m is the operating mode. It is a discrete state variable of the system and it depends of the input signal d_j (and so S_j).

Let us consider the state vector $x = [V_{C1}, V_{C2}, \dots, V_{Cp-1}, I_{Ch}]^T$. As a matter of facts, the MSC instantaneous model is defined as a state space representation such as:

$$\begin{cases} \dot{x}(t) = A_m x(t) + B_m u(t) \\ y = Cx(t), \end{cases} \tag{8}$$

where:

$$A_m = \begin{bmatrix} 0 & \cdots & 0 & \frac{d_1}{c_1} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \frac{d_{(p-1)}}{c_{(p-1)}} \\ -\frac{d_1}{L} & \cdots & -\frac{d_{(p-1)}}{L} & -\frac{R_{Ch}}{L_{Ch}} \end{bmatrix}, \tag{9}$$

$$B_m = \begin{bmatrix} 0 & \cdots & 0 & \frac{d_p}{L_{Ch}} - \frac{1}{2L_{Ch}} \end{bmatrix}^T, \tag{10}$$

$$C = [\ 0 \quad \cdots \quad 0 \quad 1\ ]. \tag{11}$$

In such a structure, $u(t) = E$, and $y(t)$ is the measure available in the MSC which corresponds to the load current.

$A_m(p \times p)$, $B_m(p \times 1)$, $C(1 \times p)$ are a variable matrices which depend on m.

The number of m is defined according to the number of cells, such as $m = 2^p$ for p cells. Subsequently, it is noticed that this model merges continuous variables (I, V_{Cj}) and discrete variables $(S_1, S_2, .., S_p)$. This is present an hybrid model of the multicell converter. To simulate the multi-cell converter with the hybrid approach every matrices A_m and B_m have to be computed. As stated in Fig. 2, the Pulse Width Modulation (PWM) provides the input order allowing to specify the operating mode m. According to m, the system switches to the appropriate matrices A and B.

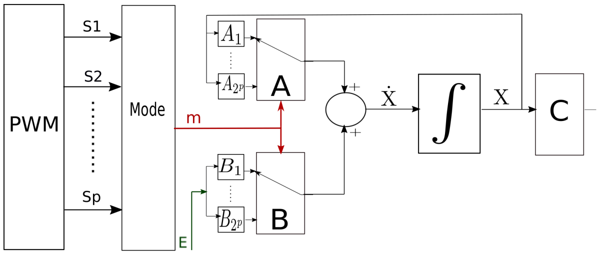


Fig. 2. Simulation model of the multicell series converter.

III. Observability Analysis

A. Hybrid Observability

Let us introduce the hybrid observability also known as the $Z(T_N)$-observability [14]. On the one hand, the hybrid time trajectory denoted T_N is a finite or infinite interval sequence $T_N = \{I_i\}_{i=0,N}$ such as:

- $I_i = [t_{i,0}, t_{i,1}]$, $I_i > 0$ for all $0 \leq i \leq N$,
- For all $0 \leq i \leq N, t_{i,1} = t_{i+1,0}$,
- $t_{0,0} = t_{init}$ and $t_{N,1} = t_{end}$.

On the other hand, $\langle T_N \rangle$ is the ordered list of the discrete input d_j associated to the trajectory T_N. It orders the operating mode applied during the interval I_i such as $\langle T_N \rangle = \{\ldots m_i \ldots\}$.

Let us define the vector $Z(t,x)$ and the projection P_i such as $P_i Z(t,x)$, is a state function of $Z(t,x)$ containing one state variable.

The $Z(T_N)$-observability consider the system (8) and a fixed hybrid time trajectory [14] T_N and $\langle T_N \rangle$. Suppose that $z = Z(t,x)$ is always continuous under any admissible control input. If there is a sequence of linear projection $\{P_i\}_{i=0,N}$ such as:

- for any $0 \leq i \leq N, P_i Z(t,x)$ is Z-observable $t \in I_i$,
- $\mathrm{rank}([\ P_0^T,\ \ \ldots\ \ ,P_N^T\ ]) = \mathrm{dim(Z)} = \mathrm{n_z}$,
- $\frac{d\bar{P}_i Z(t,x)}{dt} = 0$ for $t \in I_i$ where $\{\ \bar{P}_i^T,\ \ P_i^T\ \}$ has a full rank in $\mathbb{R}^{\mathrm{dim(z)} \times \mathrm{dim(z)}}$, then z is Z-observable with respect to the hybrid time trajectory T_N and $\langle T_N \rangle$.

The first condition indicates that a projection of the state variable $P_i Z(t,x)$ is observable, at least, on a time interval I_i of an hybrid time trajectory T_N. The second implies that all the state variables must be observable for T_N including I_i. The third condition ensures that the inobservable variable must remain constant during I_i.

B. Application of the Hybrid Observability to the MSC

Let us consider the converter's model (8), the observable vector $Z(t,x) = [V_{C1}, V_{C2}, \ldots, V_{Cp-1}]^T$ and the projection P_i such as:

$$P_i Z(t,x) = \begin{bmatrix} \delta_1 & 0 & \cdots & 0 \\ 0 & \delta_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \delta_{p-1} \end{bmatrix} \begin{bmatrix} V_{C1} \\ V_{C2} \\ \vdots \\ V_{Cp-1} \end{bmatrix}, \quad (12)$$

where for $k = \{1, \ldots, p-1\}$, $\delta_k \in \{0,1\}$. It is noticed that our wish is to observe one capacitor voltage for an interval I_i, then a single δ_k must be equal to 1.

As exemplified by [15], [16] and [17], the application of the hybrid observability yields to the following results.

The PWM strategy generates the control order d_j, $j \in \{1, \ldots, p\}$ such as each order is associated to an operating mode m_i during a time interval $I_i, i \in \{1, \ldots, N\}$. $T_N = \{I_i\}$ is the hybrid time trajectory and $\langle T_N \rangle = \{\ldots m_i \ldots\}$.

Using the following equation:

$$\dot{I}_{Ch} = -\frac{R}{L} I_{Ch} + \frac{d_p E}{L} - \frac{E}{2L} - \frac{V_{C1}}{L} d_1 - \ldots - \frac{V_{Cp-1}}{L} d_{p-1}, \quad (13)$$

it is noticed that the derivatives of the measured current allows to obtain $p-1$ independent equations that used to compute the projection P_j, such as $V_{Cj} = P_j Z(t,x)$ obtained from the measurement of I_{ch}.

Consequently $\mathrm{rank([P_1^T, P_2^T, \ldots, P_N^T])} = p - 1$.

Effectively in such control sequences d_j, the unobservable voltages are constants ($\dot{V}_{Cj} = 0$), this is verified using equation (14).

$$\dot{V}_{Cj} = I_{Ch} \frac{1}{C_j} d_j. \quad (14)$$

Consequently, the MSC is observable with respect to the hybrid time trajectory T_N and the ordered list $\langle T_N \rangle$.

IV. Observer Design

A. Observation Problem Statement

According to the previous observability study, each voltage $V_{Cj}, j = \{1, \ldots, p-1\}$ is observable according to the current measurement. As the observability is related to the mode m, the capacitor voltages are not observable through the same mode m, so it is not possible to apply a full rank observer on the hybrid model (8). The solution proposed by [18] is to break down the model into non linear sub-models where each one describes the dynamic of one voltage capacitor. Then a full rank observers were applied to the sub-models. Using this proposed solution the complexity of the estimation algorithms was increased. The purpose of our study is to develop a minimal order linear functional observers for the MSC's hybrid model in order to estimate each capacitor voltage. In fact each linear functional observer allows to estimate a linear state function of a given voltage capacitor. This section deals with the direct strategy to design a minimal order functional observer to estimate one voltage capacitor from the measurement of the load current and the knowledge of the control input.

B. Functional Observer Definition

The functional observer [19] [20] [21] [22] is a low order observer used to estimate a linear function of the state :

$$v(t) = Px(t), \quad (15)$$

where P is $\mathbb{R}^{(f \times n)}$ matrix, n is the order of the system and f is the number of rows of P. For a linear dynamic system, the linear functional observer is expressed as follow:

$$\begin{cases} \dot{z}(t) = Fz(t) + Gu(t) + Hy(t) \\ \omega(t) = Lz(t) + Vy(t). \end{cases} \quad (16)$$

where F, G, H, L and V are a constant matrices.

Let us define q as the observer order and e is the number of measures.

The objective of the study is to provide a minimal single linear observer to estimate the voltage capacitors of the MSC. On the one hand the word *minimal* indicates that the functional observer must have an order lower than $n-e-f$. Effectively, it is possible to distinguish several types of functional observers [22] according to their order. In the other hand the word single implies that $f = 1$

Moreover, in [10], a direct procedure to design this observer is presented. It is demonstrated how to guarantee a minimal q-order observer. As an hybrid system modeled in (8) is considered, an hybrid observer is proposed as:

$$\begin{cases} \dot{z}_i(t) = F_{m,i} z_{m,i}(t) + G_{m,i} u(t) + H_{m,i} y(t) \\ w_i(t) = L_{m,i} z_{m,i}(t) + V_{m,i} y(t), \end{cases} \tag{17}$$

where: m is the discrete state, i is the linear function index, $i \in \{1, \dots, s\}$, s is the number of linear functional to observe and n is the order of the system. This observer is developed to estimate the linear function $v_i(t)$ such as:

$$v_i(t) = P_i x(t), P_i \in \mathbb{R}^{(1 \times n)} \tag{18}$$

$F_{m,i}$, $G_{m,i}$, $H_{m,i}$, $L_{m,i}$ and $V_{m,i}$ are a variable matrices which depend on the operation mode m and the linear projection P_i.

The following assumptions [23] [19] [20] have to be verified to ensure that a functional observer exists:

$$G_{m,i} = T_{m,i} B_m, \tag{19}$$

$$T_{m,i} A_m - F_{m,i} T_{m,i} = H_{m,i} C_m, \tag{20}$$

$$P_i = L_{m,i} T_{m,i} + V_{m,i} C_m, \tag{21}$$

where $T_{m,i}$ is defined as $\lim_{t \to \infty} (z_i(t) - T_{m,i} x(t)) = 0$ where $e = z_i(t) - T_{m,i} x(t)$. Using (17), (19) and (20), the dynamic observation error is finally expressed as follow :

$$\dot{e}_{m,i}(t) = F_{m,i} e_{m,i}(t). \tag{22}$$

The stability of the observer is ensured by the asymptotic convergence of its error. Consquently, $F_{m,i}$ must be Hurwitz matrices.

The existence condition of a minimal order observer for a single linear functional where q is the minimal order is such as:

$$\mathrm{rank}(\Sigma_{\mathrm{q,m,i}}) = \mathrm{rank}(\begin{bmatrix} \Sigma_{q,m,i} \\ P_i A_m^q \end{bmatrix}), \tag{23}$$

where:

$$\Sigma_{q,m,i} = \begin{bmatrix} C_m \\ P_i \\ C_m A_m \\ P_i A_m \\ \vdots \\ C_m A_m^{q-1} \\ P_i A_m^{q-1} \\ C_m A_m^q \end{bmatrix}. \tag{24}$$

Consequently, $P_i A_m^q$ must be linearly dependent of the columns $C_m A_m^j$ and $P_i A^j$, $j \in \{0, \dots, q\}$. So it exists $\mathit{\Gamma}_{j,m,i}$ and $\Lambda_{j,m,i}$ for $j \in \{0, \dots, q-1\}$ such as:

$$P_i A_m^q = \sum_{j=0}^{q} \mathit{\Gamma}_{j,m,i} C_m A_m^j + \sum_{j=0}^{q-1} \Lambda_{j,m,i} P_i A_m^j. \tag{25}$$

Let us consider the linear functional v_i and its q-derivative:

$$v_i^{(q)}(t) = P_i A_m^q x(t) + \sum_{j=0}^{q-1} P_i A_m^{q-1-j} B_m u^{(j)}(t). \tag{26}$$

Using the expression (25), $v^{(q)}(t)$ can be written as follows:

$$\begin{aligned} v_i^{(q)}(t) = \sum_{j=0}^{q} \mathit{\Gamma}_{j,m,i} C_m A_m^j x(t) + \sum_{j=0}^{q-1} \Lambda_{j,m,i} P_i A_m^j x(t) + \\ \sum_{j=0}^{q-1} P_i A_m^j B_m u^{(q-1-j)}(t). \end{aligned} \tag{27}$$

In order to remove $x(t)$ from the last expression, it's possible to express $v_i^{(q)}(t)$ as follow:

$$v_i^{(q)}(t) = \sum_{j=0}^{q} \mathit{\Gamma}_{j,m,i} y^{(i)} + \sum_{j=0}^{q-1} \Lambda_{j,m,i} v^{(j)} + \sum_{j=0}^{q-1} \phi_{j,m,i} u^{(j)}(t), \tag{28}$$

where $k \in \{0, \dots, q-2\}$ and Φ_k is expressed as follow:

$$\begin{aligned} \phi_k = & \left[P_i A_m^{q-1-k} - \sum_{j=k+1}^{q} \mathit{\Gamma}_{j,m,i} C_m A_m^{j-k-1} \right] B_m \\ & - \left[\sum_{j=k+1}^{q-1} \Lambda_{j,m,i} P_i A_m^{j-k-1} \right] B_m, \end{aligned} \tag{29}$$

and

$$\phi_{q-1} = \left[P_i - \mathit{\Gamma}_{q,m,i} C_m \right] B_m.$$

Thus the observer matrices are presented in expression (30).

$$\begin{aligned} \dot{z}_{m,i}(t) = & \begin{bmatrix} 0 & & 0 & \Lambda_{0,m,i} \\ 1 & \ddots & & \Lambda_{1,m,i} \\ 0 & \ddots & 0 & \vdots \\ & & 1 & \Lambda_{q-1,m,i} \end{bmatrix} z(t) + \begin{bmatrix} \phi_{0,m,i} \\ \phi_{1,m,i} \\ \vdots \\ \phi_{q-1,m,i} \end{bmatrix} u(t) \\ & + \begin{bmatrix} \Gamma_{0,m,i} + \Lambda_{0,m,i} \Gamma_{q,m,i} \\ \Gamma_{1,m,i} + \Lambda_{1,m,i} \Gamma_{q,m,i} \\ \vdots \\ \Gamma_{q-1,m,i} + \Lambda_{q-1,m,i} \Gamma_{q,m,i} \end{bmatrix} y(t) \end{aligned} \tag{30}$$

$$\omega_{m,i}(t) = \begin{bmatrix} 0 & \dots & 0 & 1 \end{bmatrix} z_{m,i}(t) + \mathit{\Gamma}_{q,m,i} y(t).$$

Note that a rigorous proof of this results for a dynamic system is established in [10]. According to this result it is possible to deduce the previous expression of an hybrid functional observer applied for each operation mode of the MSC. The parameters $\Gamma_{j,m,i}$ and $\Lambda_{j,m,i}$ used to define the observer matrices $F_{m,i}, G_{m,i}, H_{m,i}, L_{m,i}$ and $V_{m,i}$ are deduced. Finally, from expression (25), the observer gains can be obtained as follow:

$$P_i A_m \Sigma^{\dagger}_{q,m,i} = \\ [\ \Gamma_{0,m,i} \quad \Lambda_{0,m,i} \quad \Gamma_{1,m,i} \quad \ldots \quad \Lambda_{q-1,m,i} \quad \Gamma_{q,m,i}\ ], \tag{31}$$

where $\Sigma^{\dagger}_{q,m,i}$ is the pseudo inverse of $\Sigma_{q,m,i}$.

C. Z(TN) Observability Applied to the MSC Hybrid Model

The second condition of the $Z(TN)$-observability of the MSC states that: $\mathrm{rank}([|\mathrm{P}_1|, |\mathrm{P}_2|, \ldots, |\mathrm{P_N}|]) = \mathrm{p}-1$ such as if $P_i Z(t, x)$ is observable then, using expression (12), $|\delta_i| = 1$ and for $k \neq i$, $|\delta_k| = 0$.

To fulfill this condition the PWM control should be able to generate an order control which allow to have the elementary sequences $(1, 0, 0, \ldots, 0)$, ..., $(0, 0, 0, \ldots, 1)$, hence the idea of using the following linear functions:

$$P_1 = [\ 1 \quad 0 \quad \ldots \quad 0 \quad 0\ ], P_2 = [\ 0 \quad 1 \quad 0 \quad \ldots \quad 0\ ], \\ \ldots, P_{p-1} = [\ 0 \quad \ldots \quad 0 \quad 1 \quad 0\ ]. \tag{32}$$

To conclude, it means that $Z(T_N)$ observability allows stating that P_{p-1} projection matrices exists for V_{Cp-1} capacitor voltage observers. The design of minimal linear functional observers applied to hybrid system leads that for every P_{p-1} and for every operating mode it exists an observer as stated in (17) that can be expressed. Finally, as the $Z(T_N)$-observability is demonstrated for the MSC hybrid system in Section III-B, a battery of observer designed from the operating modes and per capacitor voltage, will be used for the hybrid trajectory $\{T_N, < T_N >\}$.

The simulation model of the functional observer for MSC is shown in Fig. 3. The next section is dedicated the design of a minimal functional observer for a four cells MSC.

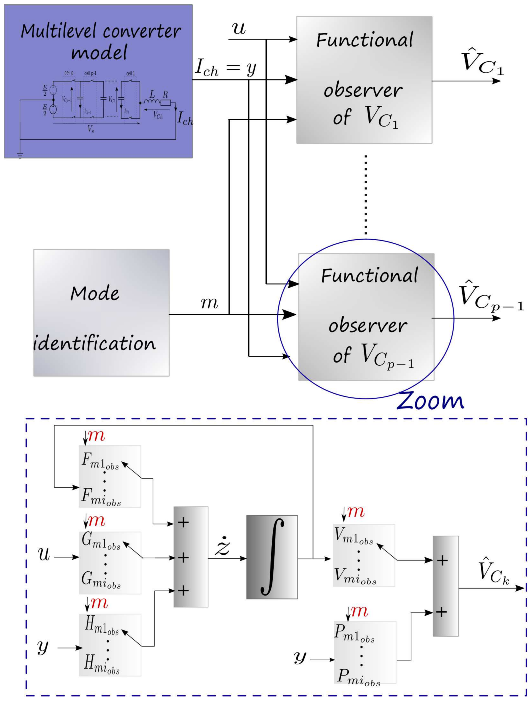


Fig. 3. Simulation model of the functional observer for MSC

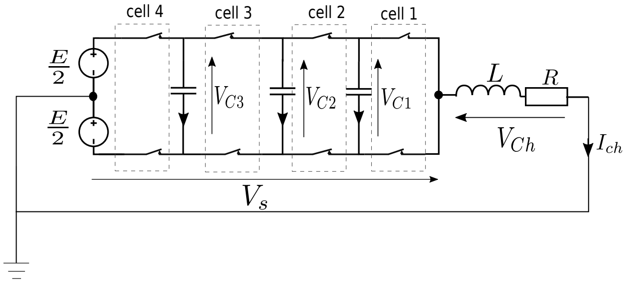


Fig. 4. Four-cells multicell series converter

V. Example of Four-Cells MSC

Let us consider a four-cells MSC presented in Fig. 4. Sixteen possible modes of operation can be found.

Table I summarizes the characteristics of each operation mode of four-cells MSC. The level of the output voltage reached is also indicated. $dj = S_{(j+1)} - S - j$ and $d_p = S_p$ are defined knowing at each mode the switch state S_j.

It is noticed that $x(t) = [V_{C1}, V_{C2}, V_{C3}, I_{ch}]^T \in \mathbb{R}^4$ is the state vector. Then, the hybrid model of the four-cells MSC is presented as follow:

$$A_m = \begin{bmatrix} 0 & 0 & 0 & \frac{d_1}{C} \\ 0 & 0 & 0 & \frac{d_2}{C} \\ 0 & 0 & 0 & \frac{d_3}{C} \\ -\frac{d_1}{L_{ch}} & -\frac{d_2}{L_{ch}} & -\frac{d_3}{L_{ch}} & -\frac{R_{Ch}}{L_{Ch}} \end{bmatrix}, \\ B_m = \begin{bmatrix} 0 & 0 & 0 & \frac{d_p}{L_{Ch}} - \frac{1}{2L_{Ch}} \end{bmatrix}, \\ C_m = [\ 0 \quad 0 \quad 0 \quad 1\ ]. \tag{33}$$

In order to check the accuracy of the proposed model, some simulations are done. The parameters used are given in Table II.

TABLE I. Operations modes for p=4

Mode (m)	d_4	d_3	d_2	d_1	V_s
1	0	0	0	0	0
2	0	0	0	-1	$V_s = V_{C1}$
3	0	0	-1	1	$V_s = V_{C2} - V_{C1}$
4	0	0	-1	0	$V_s = V_{C2}$
5	0	-1	1	0	$V_s = V_{C3} - V_{C2}$
6	0	-1	1	-1	$V_s = V_{C3} - V_{C2} + V_{C1}$
7	0	-1	0	1	$V_s = V_{C3} - V_{C1}$
8	0	-1	0	0	$V_s = V_{C3}$
9	1	1	0	0	$V_s = E - V_{C3}$
10	1	1	0	-1	$V_s = E - V_{C3} + V_{C1}$
11	1	1	-1	1	$V_s = E - V_{C3} + V_{C2} - V_{C1}$
12	1	1	-1	0	$V_s = E - V_{C3} + V_{C2}$
13	1	0	1	0	$V_s = E - V_{C2}$
14	1	0	1	-1	$V_s = E - V_{C2} + V_{C1}$
15	1	0	0	1	$V_s = E - V_{C1}$
16	1	0	0	0	E

TABLE II. Simulation parameters

E	f_{PWM}	$f_{modulante}$	L_{ch}	R	C_1, C_2, C_3
$230\ V$	$1\,kHz$	$50\ Hz$	$1\,mH$	10Ω	$4 \times 10^{-4}\ F$

The simulation results allows to get the load voltage and the load current which are presented, respectively in Fig. 5 and Fig. 6. A sinus PWM is used with triangle carrier based signals.

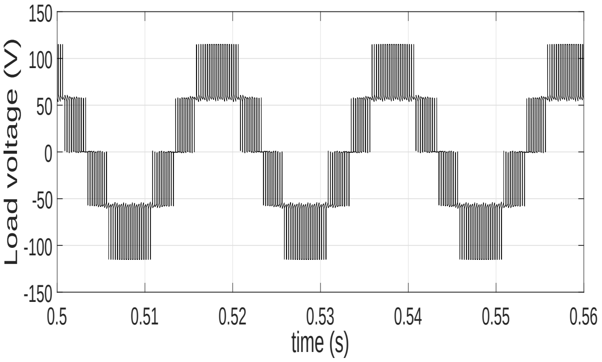


Fig. 5. Load voltage

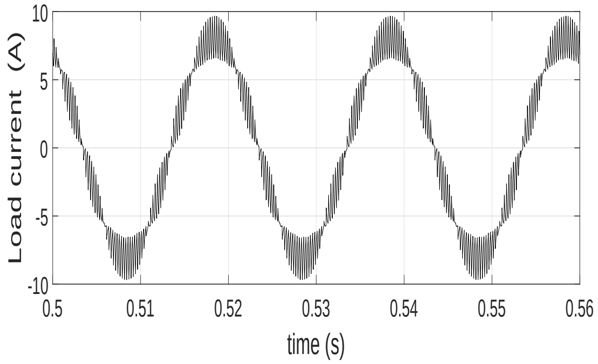


Fig. 6. Load current

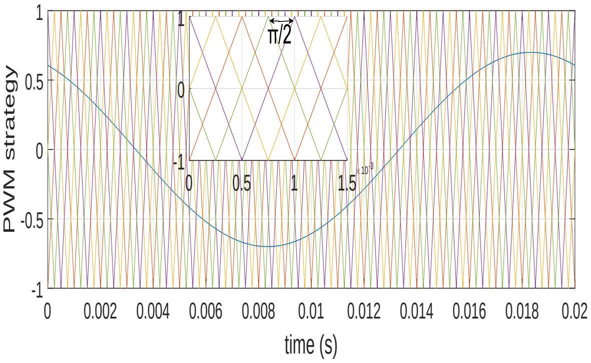


Fig. 7. PWM strategy

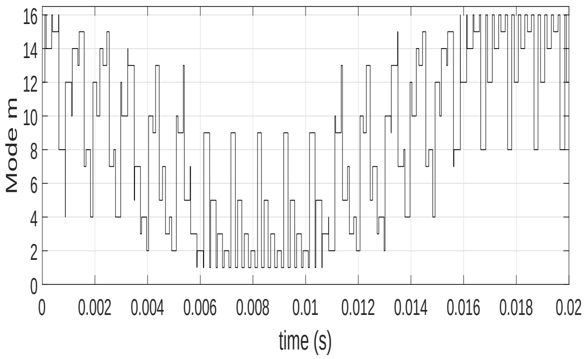


Fig. 8. Switching signal

In this control strategy, there are four triangular carrier waves with a phase shift of $\frac{2\pi}{p} = \frac{\pi}{2}$ for each switching cell. The control strategy and the operating mode value are depicted in Fig. 7 and 8 for a one period $\frac{1}{f} = 0.02s$.

Fig. 7 and 8 shows that for a modulating time period, we pass through all the operating mode m.

A. Design of Minimal Functional Observer

The following steps illustrate the procedure to get the minimal functional observers in order to estimate V_{C1}. In this case $i = 1$, so:

$$P_1 = [\ 1 \quad 0 \quad 0 \quad 0\ ]. \tag{34}$$

It's noticed that this procedure is the same that can be used to estimate the other voltages of the MSC.

1) Mode classification estimation approach: Our proposed observation approach is based on the $Z(TN)$-observability criterion. In fact, the $Z(TN)$-observability analysis prove the existence of a time interval I_i associated to an operating mode through it the voltage V_{C1} is observable. During this interval the other voltages (V_{C2} and V_{C3}) are unobservable and keep constant. Also, it exist an operating modes in which V_{C1} is unobservable and constant. For other modes V_{C1} is not able to be estimated but it evolves over time, in this case there an estimation of a sum of voltage V_{Cj} which caused a V_{C1} voltage information gathering. Our proposed observation approach is based on this mode classification. Therefore, the first functional observers are designed for the operating modes that allows the estimation of V_{C1} (modes 2 and 15). The second observers are designed for the modes in which V_{C1} is unobservable but evolves over time (modes 3, 6, 7, 10, 11, 14). Finally, for the last class of modes, we set zero dynamics to the estimated voltage ($\dot{V}_{C1} = 0$). This observation approach is presented in Fig. 9. The first step is to define the minimal order q by using the

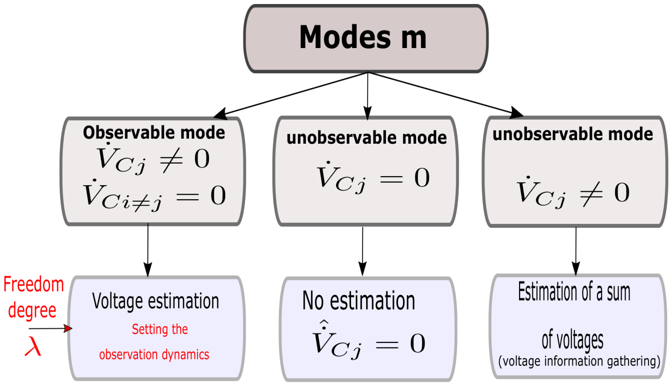


Fig. 9. Observation approach based on mode classification

existence condition (23).

2) Operation modes 2 *and* 15 *(Observable modes of* V_{C1}*):* Let us consider the observable modes 2 and 15 of V_{C1}, for $q = 1$:

$$\Sigma_{1,2,1} = \begin{bmatrix} C \\ P_1 \\ CA_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1000 & 0 & 0 & -10000 \end{bmatrix}, \quad (35)$$

$$\Sigma_{1,15,1} = \begin{bmatrix} C \\ P_1 \\ CA_{15} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ -1000 & 0 & 0 & -10000 \end{bmatrix}, \quad (36)$$

As $P_1A_2 = \begin{bmatrix} 0 & 0 & 0 & -2500 \end{bmatrix}$ and $P_1A_{15} = \begin{bmatrix} 0 & 0 & 0 & 2500 \end{bmatrix}$,
then

$$\text{rank}(\Sigma_{1,2,1}) = \text{rank}(\begin{bmatrix} \Sigma_{1,2,1} \\ P_1A_2 \end{bmatrix}) = 2 \quad (37)$$

and

$$\text{rank}(\Sigma_{1,15,1}) = \text{rank}(\begin{bmatrix} \Sigma_{1,15,1} \\ P_1A_{15} \end{bmatrix}) = 2 \quad (38)$$

Consequently, a minimum observer can be designed to estimate the voltage V_{C1} of the four-cells MSC.

As $\Sigma_{1,m,1}$ is singular and $\dim(\ker(\Sigma_{1,m,1})) = 1$, for $m = \{2, 15\}$, the observer gains can be expressed as follow [24]:

$$\begin{aligned} \begin{bmatrix} \mathit{\Gamma}_{0,m,1} & \Lambda_{0,m,1} & \Gamma_{1,m,1} \end{bmatrix} &= P_1A_m\mathit{\Sigma}^{\dagger}_{1,m,1} + \\ & Z(I_3 - \mathit{\Sigma}^{\dagger}_{1,m,1}\mathit{\Sigma}_{1,m,1}), \end{aligned} \quad (39)$$

where $\mathit{\Sigma}^{\dagger}_{1,m,1}$ is the pseudo inverse matrix of $\mathit{\Sigma}_{1,m,1}$ and I_3 is the (3×3) identity matrix and Z an arbitrary matrix.

Let us consider the scalar $\lambda_m = Z\phi$ is the freedom degree such as ϕX is the full rank factorization of $(I_3 - \mathit{\Sigma}^{\dagger}_{1,m,1}\mathit{\Sigma}_{1,m,1})$, so the gains observer are expressed as follow:

$$\begin{bmatrix} \mathit{\Gamma}_{0,m,1} & \Lambda_{0,m,1} & \Gamma_{1,m,1} \end{bmatrix} = P_1A_m\mathit{\Sigma}^{\dagger}_{1,m,1} + \lambda_m X. \quad (40)$$

For $m = \{2, 15\}$, λ_m is set such that $F_{m,1}$ is a Hurwitz matrix. For this modes, the freedom degree allows to fix the observer dynamic. For $m = 2$ the gains observer are expressed as follow:

$$\begin{aligned} \mathit{\Gamma}_{0,2,1} &= 24.7525 - 0.995\lambda_2, \\ \Lambda_{0,2,1} &= -247.5248 + 0.0995\lambda_2, \\ \Gamma_{1,2,1} &= -0.2475 - 0.0001\lambda_2. \end{aligned} \quad (41)$$

Based on expressions (30) and (42), $F_{2,1}$ is a Hurwitz matrix if $\lambda_2 < 2487.686$.

For $m = 15$, the gains observer are expressed as follow:

$$\begin{aligned} \mathit{\Gamma}_{0,15,1} &= 24.7525 - 0.995\lambda_{15}, \\ \Lambda_{0,15,1} &= -247.5248 - 0.0995\lambda_{15}, \\ \Gamma_{1,15,1} &= -0.2475 - 0.0001\lambda_{15}. \end{aligned} \quad (42)$$

So, $F_{15,1}$ is a Hurwitz matrix if $\lambda_{15} > -2487.686$.

Based on observer matrices expressions (30), we obtain the observer gains for mode 2 and 15 which are illustrated in Table III. The obtained pole is -98.2 which indicate the stability of the minimal functional observers.

TABLE III. OBSERVER GAINS FOR MODES 2 AND 15

m	λ_m	$F_{m,1}$	$G_{m,1}$	$H_{m,1}$	$L_{m,1}$	$V_{m,1}$
2	1500	-98.2	49.13	-1527	1	0.09
15	-1500	-98.2	49.13	1527	1	-0.09

3) Case of modes 3, 6, 7, 10, 11 *and* 14 *:* Let us consider the modes 3, 6, 7, 10, 11 and 14 in which V_{C1} is unobservable but it is not constant. We test the existence condition for $q = 1$ and for the mode $m = 3$.

$$\Sigma_{1,3,1} = \begin{bmatrix} C \\ P_1 \\ CA_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ -1000 & 1000 & 0 & -10000 \end{bmatrix}. \quad (43)$$

As

$$P_1A_3 = \begin{bmatrix} 0 & 0 & 0 & 2500 \end{bmatrix}, \quad (44)$$

then

$$\text{rank}(\Sigma_{1,3,1}) = \text{rank}(\begin{bmatrix} \Sigma_{1,3,1} \\ P_1A_3 \end{bmatrix}) = 3. \quad (45)$$

This existence condition is also true for the operating modes 6, 7, 10, 11, 14. Consequently, a minimum observer can be designed for this operation modes of the MSC.

As $\dim(\ker(\Sigma_{1,m,1})) = 0$, for $m = 3, 6, 7, 10, 11, 14$, there is no freedom degree, so the observer gains are expressed as follow:

$$\begin{bmatrix} \mathit{\Gamma}_{0,m,1} & \Lambda_{0,m,1} & \Gamma_{1,m,1} \end{bmatrix} = P_1A_m\mathit{\Sigma}^{\dagger}_{1,m,1}. \quad (46)$$

Using the gains expression from (46), and based on a the expression (30) of the functional observer, the obtained gains observer values for this modes are presented in Table IV. From Table III and IV we deduce that these operation modes are not intervene on observation dynamic of the voltage V_{C1}. In fact, the observation dynamics and the convergence are defined through the modes 2 and 15 in which the output Ich depend only on the voltage V_{C1} thanks to the freedom degree. The same observer design strategy are used for the estimation of the voltage V_{C2} and V_{C3}. The obtained gains observer values are given in the appendix.

TABLE IV. OBSERVER GAINS FOR MODES 3, 6 ,7, 10, 11, 14

m	$F_{m,1}$	$G_{m,1}$	$H_{m,1}$	$L_{m,1}$	$V_{m,1}$
3	0	0	2500	1	0
6	0	0	-2500	1	0
7	0	0	2500	1	0
10	0	0	-2500	1	0
11	0	0	2500	1	0
14	0	0	-2500	1	0

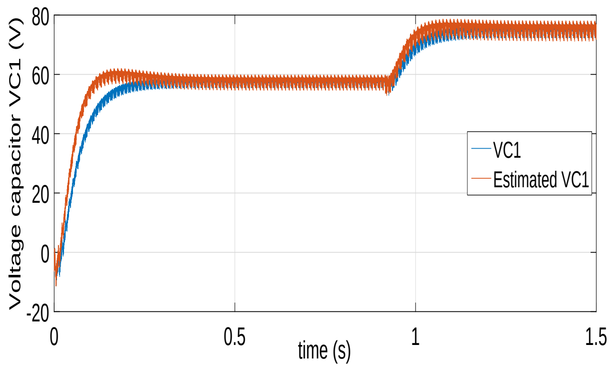


Fig. 10. Estimation of voltage capacitor C1

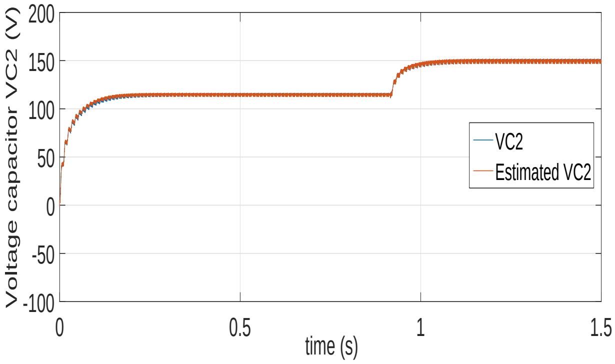


Fig. 11. Estimation of voltage capacitor C2

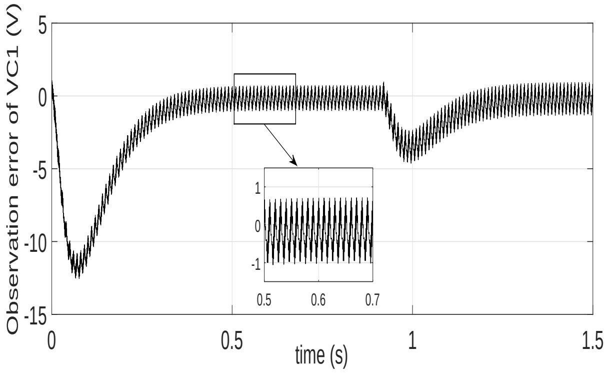


Fig. 13. Observation error of voltage capacitor C1

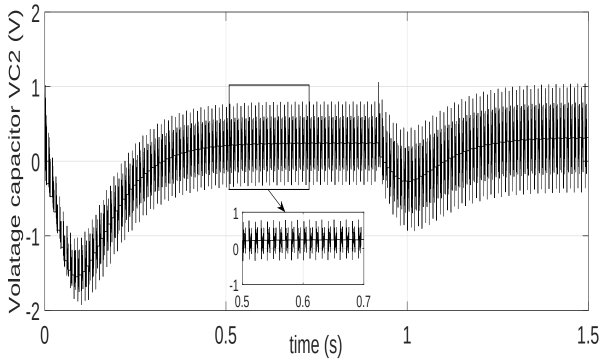


Fig. 14. Observation error of voltage capacitor C2

B. Simulation Results

The simulation results showing the performance of the minimal functional observer for the all voltages capacitor are presented in Fig. 10, 11, 12, 13, 14 and 15.

One can note that this observer keeps good properties in term of asymptotic convergence. In fact for $E = 230V$, the estimation of the voltage capacitors converge to its expected steady state values which are $V_{C1} = 57.5V$, $V_{C2} = 115V$ and $V_{C3} = 172.5V$. In order to test the robustness of this observer, at $t = 0.8s$ the input voltage is set to $300V$. The estimated voltage values respond to the abrupt change of the input signal.

C. Discussion

As presented in Fig. 16, 17 and 18, the estimated capacitor voltage V_{Cj} is variable when the system switch to an observable mode, and keep constant when an other voltage is observable to avoid its re-observation. This result prove the respect of $Z(TN)$-observability criterion.

The estimation approach adopted in this paper are based on the $Z(TN)$-observability. The observer proposed in this paper

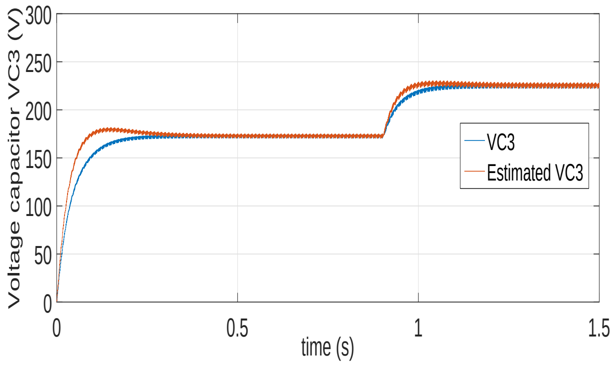


Fig. 12. Estimation of voltage capacitor C3

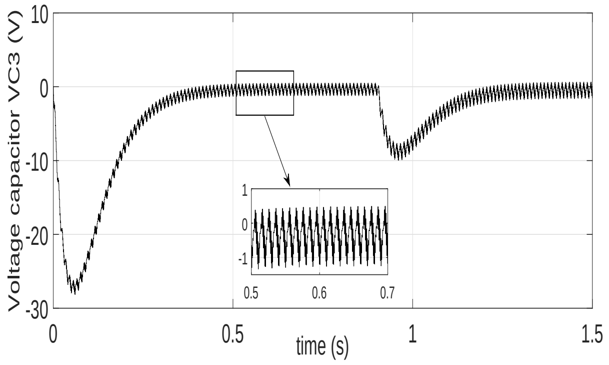


Fig. 15. Observation error of voltage capacitor C3

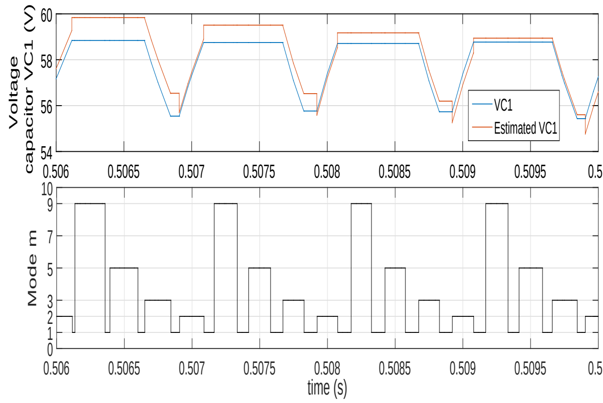


Fig. 16. Zoom on voltage capacitor C1 and its estimation

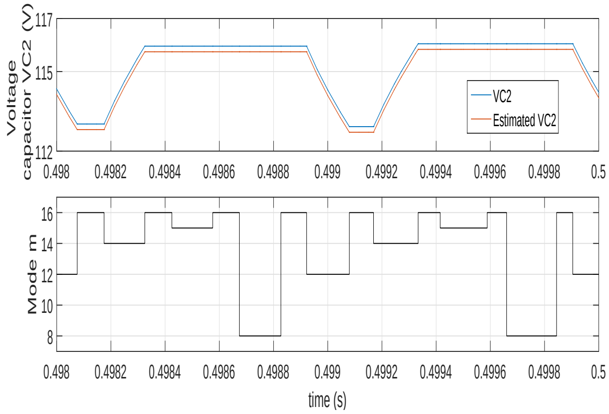


Fig. 17. Zoom on voltage capacitor C2 and its estimation

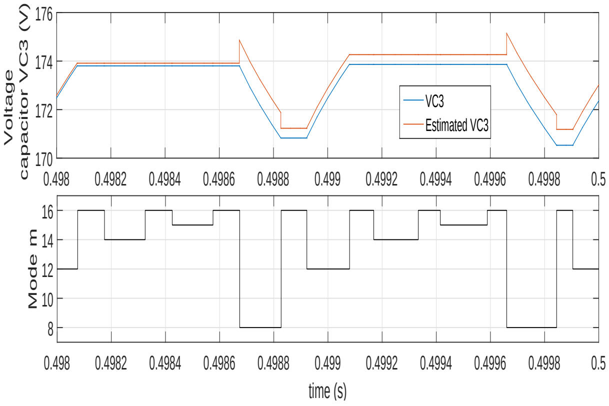


Fig. 18. Zoom on voltage capacitor C3 and its estimation

is the minimal linear functional observer. On one hand this observer aims at reducing the system order and estimating each voltage separately on the other hand. A functional observer battery is designed for each voltage and for different operating modes. The adopted strategy in the functional observer design for each voltage capacitor is based on the classification of the operating modes which is a way to simplify the observation algorithm. In fact, When the system switch to the modes through it the voltage is not observable and constant, a zero dynamic is set. We deduce that the classification allows to specify the modes which intervene in the observation and neglects the other modes. As a results, the operating modes used for the voltage estimation are reduced from 16 to 8 modes (for four-cells converter). The simulation result prove that only the modes in which the voltage is observable, intervene on the observation dynamic.

This method becomes more complicated as the number of cells increases. In this case a huge number of battery of linear function observers are needed to estimate each voltages of the floating capacitors, there are also a significant number of modes.

VI. Conclusion

In this paper the minimal single functional observer is proposed in order to estimate the capacitor voltages of the multicell converter which belongs to the hybrid dynamic system class. The problem of the capacitor voltages estimation is solved. Some simulations results of the four cells converters illustrate the performance and robustness of the proposed observer in the dynamic behaviour. This software sensor is an effective solution to detect the voltage imbalance instead of extra sensors. The deterioration of floating capacitors can be the cause of voltage unbalance. Consequently, our further work involves the application of the observer introduced in this paper in the presence of a capacitor default. Later, a synthesis of a closed-loop control using the estimated voltages in order to compensate this default is feasible .

VII. Appendix

A. Estimation of V_{C2}

To estimate the voltage capacitor V_{C2}, the vector P_2 is set such as:

$$P_2 = [\ 0 \quad 1 \quad 0 \quad 0\ ]. \tag{47}$$

The observable mode of V_{C2} are modes 3, 4, 5, 6, 11, 12, 13, 14. The modes liable to the observation dynamic of the voltage V_{C2} are the modes 4 and 13. So, the observer gains are presented in Table V. $\lambda_4 = 1500$ and $\lambda_{13} = -1500$.

TABLE V. Observer gains for modes 3, 4 ,5, 6, 11, 12, 13, 14

m	$F_{m,1}$	$G_{m,1}$	$H_{m,1}$	$L_{m,1}$	$V_{m,1}$
3	0	0	-2500	1	0
4	-98.26	49.13	-1.527	1	0.0983
5	0	0	2500	1	0
6	0	0	2500	1	0
11	0	0	-2500	1	0
12	0	0	-2500	1	0
13	-98.26	49.13	1527	1	-0.0983
14	0	0	2500	1	0

B. Estimation of V_{C3}

To estimate the voltage capacitor V_{C3}, the vector P_3 is set such as:

$$P_3 = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}. \tag{48}$$

Observable mode of V_{C3} are modes 5, 6,7, 8, 9, 10, 11, 12. The modes liable to the observation dynamic of the voltage V_{C3} are the modes 8 and 9. So, The observer gains are presented in Table VI.

TABLE VI. Observer gains for modes 5, 6, 7, 8, 9, 10, 11, 12

m	$F_{m,1}$	$G_{m,1}$	$H_{m,1}$	$L_{m,1}$	$V_{m,1}$
5	0	0	-2500	1	0
6	0	0	-2500	1	0
7	0	0	-2500	1	0
8	-98.26	49.13	-1527	1	0.0983
9	-98.26	49.13	1527	1	-0.0983
10	0	0	2500	1	0
11	0	0	2500	1	0
12	0	0	2500	1	0

$\lambda_8 = 1500$ and $\lambda_9 = -1500$.

References

[1] M. Kasper, D. Bortis, and J. W. Kolar, "Scaling and balancing of multi-cell converters," in *2014 International Power Electronics Conference (IPEC-Hiroshima 2014-ECCE ASIA)*. IEEE, 2014, pp. 2079–2086.

[2] M. Jday and J. Haggège, "Modeling and neural networks based control of power converters associated with a wind turbine," in *2017 International Conference on Green Energy Conversion Systems (GECS)*. IEEE, 2017, pp. 1–7.

[3] A. K. Sadigh, V. Dargahi, and K. A. Corzine, "New active capacitor voltage balancing method for flying capacitor multicell converter based on logic-form-equations," *IEEE Transactions on industrial electronics*, vol. 64, no. 5, pp. 3467–3478, 2016.

[4] N. Gazzam and A. Benalia, "Observability analysis and observer design of multicellular converters," in *2016 8th International Conference on Modelling, Identification and Control (ICMIC)*. IEEE, 2016, pp. 763–767.

[5] J. Van Gorp, M. Defoort, M. Djemai, and K. C. Veluvolu, "Fault detection based on higher-order sliding mode observer for a class of switched linear systems," *IET Control Theory & Applications*, vol. 9, no. 15, pp. 2249–2256, 2015.

[6] K. Benmansour, J. De Leon, and M. Djemai, "Adaptive observer for multi-cell chopper," in *Second International Symposium on Communications, Control and Signal Processing ISCCSP, Marrakech, Maroc*, 2006.

[7] M. Ghanes, F. Bejarano, and J.-P. Barbot, "On sliding mode and adaptive observers design for multicell converter," in *2009 American Control Conference*. IEEE, 2009, pp. 2134–2139.

[8] M. Jday, P.-E. Vidal, J. Haggège, and F. Rotella, "Observability and sliding mode observer design for multi-cell series converter," in *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*. IEEE, 2019, pp. 1486–1491.

[9] J. Lin, Y. Shi, Z. Gao, and J. Ding, "Functional observer for switched discrete-time singular systems with time delays and unknown inputs," *IET Control Theory & Applications*, vol. 9, no. 14, pp. 2146–2156, 2015.

[10] I. Sakhraoui, B. Trajin, and F. Rotella, "Application of linear functional observers for the thermal estimation in power modules," 2018.

[11] K. Berkoune, P.-E. Vidal, and F. Rotella, "Modélisation générique pour les stratégies de modulation des onduleurs multiniveaux: application aux onduleurs à capacités flottantes," 2016.

[12] J. Rodriguez, J.-S. Lai, and F. Z. Peng, "Multilevel inverters: a survey of topologies, controls, and applications," *IEEE Transactions on industrial electronics*, vol. 49, no. 4, pp. 724–738, 2002.

[13] A. Bouarfa, M. Bodson, and M. Fadel, "A fast active-balancing method for the 3-phase multilevel flying capacitor inverter derived from control allocation theory," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 2113–2118, 2017.

[14] W. Kang and J.-P. Barbot, "Discussions on observability and invertibility," *IFAC Proceedings Volumes*, vol. 40, no. 12, pp. 426–431, 2007.

[15] B. Amghar, A. Darcherif, and J.-P. Barbot, "Z (tn)-observability and control of parallel multicell chopper using petri nets," *IET Power Electronics*, vol. 6, no. 4, pp. 710–720, 2013.

[16] A. Smati, W. Chagra, and M. Kssouri, "Fault-tolerant model predictive control for az (tn)-observable linear switching systems," *International Journal Of Advanced Computer Science And Applications*, vol. 8, no. 6, pp. 365–374, 2017.

[17] F. J. Bejarano, M. Ghanes, and J.-P. Barbot, "Observability and observer design for hybrid multicell choppers," *International Journal of Control*, vol. 83, no. 3, pp. 617–632, 2010.

[18] J. Van Gorp, M. Defoort, M. Djemai, and N. Manamanni, "Hybrid observer for the multicellular converter," *IFAC Proceedings Volumes*, vol. 45, no. 9, pp. 259–264, 2012.

[19] M. Darouach, "Existence and design of functional observers for linear systems," *IEEE Transactions on Automatic Control*, vol. 45, no. 5, pp. 940–943, 2000.

[20] H. Trinh, T. Tran, and S. Nahavandi, "Design of scalar functional observers of order less than (ν- 1)," *International journal of control*, vol. 79, no. 12, pp. 1654–1659, 2006.

[21] H. Trinh and T. Fernando, "On the existence and design of functional observers for linear systems," in *2007 International Conference on Mechatronics and Automation*. IEEE, 2007, pp. 1974–1979.

[22] T. L. Fernando, H. M. Trinh, and L. Jennings, "Functional observability and the design of minimum order linear functional observers," *IEEE Transactions on Automatic Control*, vol. 55, no. 5, pp. 1268–1273, 2010.

[23] F. Rotella and I. Zambettakis, "A note on functional observability," *IEEE Transactions on Automatic Control*, vol. 61, no. 10, pp. 3197–3202, 2015.

[24] A. Ben-Israel and T. N. Greville, *Generalized inverses: theory and applications*. Springer Science & Business Media, 2003, vol. 15.

Binning Approach based on Classical Clustering for Type 2 Diabetes Diagnosis

Hai Thanh Nguyen[1]
College of Information and
Communication Technology
Can Tho University
Can Tho, Vietnam

Nhi Yen Kim Phan[2]
College of Information and
Communication Technology
Can Tho University
Can Tho, Vietnam

Huong Hoang Luong[3]
Department of Information Technology
FPT University
Can Tho, Vietnam

Nga Hong Cao[4]
Department of Computer Science and
Information Engineering
National Central University
Taiwan

Hiep Xuan Huynh[5]
College of Information and
Communication Technology
Can Tho University
Can Tho, Vietnam

***Abstract*—In recent years, numerous studies have been focusing on metagenomic data to improve the ability of human disease prediction. Although we face the complexity of disease, some proposed frameworks reveal promising performances in using metagenomic data to predict disease. Type 2 diabetes (T2D) diagnosis by metagenomic data is one of the challenging tasks compared to other diseases. The prediction performances for T2D usually reveal poor results which are around 65% in accuracy in state-of-the-art. In this study, we propose a method combining K-means clustering algorithm and unsupervised binning approaches to improve the performance in metagenome-based disease prediction. We illustrate by experiments on metagenomic datasets related to Type 2 Diabetes that the proposed method embedded clusters generated by K-means allows to increase the performance in prediction accuracy reaching approximately or more than 70%.**

***Keywords*—Unsupervised binning; K-means clustering algorithm; metagenomics; metagenome-based disease prediction; Type 2 diabetes diagnosis**

I. Introduction

Metagenomics (Environmental Genomics, Ecogenomics or Community Genomics) is the study of genetic material recovered directly from environmental samples. Metagenomics is directly the study of communities of microbial organisms in their natural environments by applying modern genomic techniques that pass the need for isolation and lab cultivation of individual species [1], [2], [3], [4], [5], [6]. Reassembly of multiple genomes has provided insight into energy and nutrient cycling within the community, genome structure, gene function, population genetics and microheterogeneity, and lateral gene transfer among members of an uncultured community. The application of metagenomic sequence information will facilitate the design of better culturing strategies to link genomic analysis with pure culture studies. Why do we study metagenomics? As in [2] mentioned that Metagenomics has brought us discovery of novel natural products, new antibiotica, new molecules with new functions, new enzymes and bioactive molecules, what is a genome or species, diversity of life, interplay between human and microbes, how do microbial communities work and how stable are they, holistic view on biology. Metagenomics cloned specific gene sequences (usually 16S rRNA genes) to conduct data on the biodiversity of environmental samples. With traditional genetic and microbiological studies of genomes sequencing of microorganisms based on cultured lineage samples, it was found that it would be impossible to biodiversity of microorganisms. Therefore, metagenomics plays an important role in helping humans discover microbial diversity. In medicine, the microbial community plays a very important role in protecting human health. Therefore, the purpose of metagenomics is to understand the composition and activity of complex microbial groups in environmental samples through analysis of their DNA sequences. On the other hand, there are numerous data on multiple genomes that we can carry out a series of gene isolation projects depending on the purpose of the research.

Metagenomic is an improved method compared to traditional microbiology, the research of metagenomes obtained from genetic material from first samples, without the need for laboratory cultures. This method is commonly used on the human intestine because it is the place where the digestive process, metabolism and has 10 times the total number of cells of the body. Based on metagenomics, we can develop algorithms to predict disease, determine a patient's sensitivity and then offer reasonable treatments. However, the disease is complicated in diagnosis and prognosis and we only have a limited amount of data to observe.

Type 2 diabetes (T2D) is a heterogeneous metabolic disorder that damages many organs of the body. The disease tends to increase due to the influence of modern life, bad living habits. Nowadays, the prediction is not highly accurate and the treatment is commonly applied to patients diagnosed with some similar manifestations. With that treatment, we find that genetic diversity has not been effectively applied, leading to an improvement in the health of some patients. The performances on models for predicting T2D usually yield poor results.

II. Related Work

As mentioned above, metagenomics is an approach that utilizes extraction of genomic information directly from the environmental sample. So that, genetic information samples are more representative for a given environment and supplies a better insight into microbial environmental and metabolic diversity. By using next-generation sequencing in metagenomics project to determine genetic potential in microbial communities from a wealth of environmental niches, including those linked with human body and relative with human healthcare. Human microbiome in health and disease plays a significant role that has recently been given considerable observation [7], and distinct diseases have been associated with gut microbiota [7], [8], [9], [10], [11], [12], [13], [14], [15]. With respect to, experience 's Maja and et al [8] that a bias in codon usage present throughout the entire microbial community by applying definitions of translational optimization through codon usage adaptation on completely metagenomic datasets. They can be used as a powerful analytical tool for predicting community lifestyle-specific metabolism. Moreover, Maja and et al demonstrate this approach combined with machine learning, to classify microbiome samples in human gut according to the pathological condition diagnosed in the human host. In addition, predicting disease-relevant features in microbial gut metagenomes by using the principle of utilizing the prokaryotic translational optimization effect combined with the machine learning based classification and enriched gene datasets that explore a supportive method to analyzing metagenomic datasets. Authors in [8], [16] proposed methods using machine learning and deep learning to do disease prediction tasks and obtained promising results.

K-means clustering is an unsupervised learning algorithm. From the input data without the label to be clustered and the number of clusters to be divided, we will use the algorithm to divide the data into clusters of similar properties. Applications of clustering algorithms have been used commonly to resolve data clustering. Based on clustering methods, we can obtain a meaningful intuition of the structure of the data. Moreover, we can use "Cluster-then-predict". That means, we observe generated clusters, then different models will be built for various subgroups if there exists a wide variation in the behaviors of a variety of subgroups. Numerous studies in biological computation tasks have been applying k-mean to do specific analyses. Authors in [17] used k-mean to process Microarray data for bioinformatics tasks. [18] also implemented k-mean to cluster biological sequences by first converting them into an intermediate binary format where Hamming distance is used as the metric of comparison. The research in [19] presented enhanced k-mean to do Bioinformatics Data Clustering. In 2019, a study [20] introduce a modified sparse K-means clustering method to detect risk genes involved with Type II Diabetes Mellitus. From some previous results, we can see potential benefits to leverage k-mean in bioinformatics tasks.

In recent years, the application of machine learning algorithms to study metagenomic has become popular and the accuracy of diagnosis has been improved over time. In this article, we propose the application of the K-means clustering algorithm in the binning approach to improve the accurate results in predicting T2D. We leverage k-mean clustering as a tool to support binning data. By identifying clusters which can exist in the data, we hope to improve the performance via using a binning approach. Our study's contribution is multi-fold:

- We present results of various binning approaches on Type Diabetes disease using metagenomic data which appear as a very big challenge for diagnosis.
- The work aims to illustrate a potential advantage of using clustering algorithms to identify breaks for binning approaches to obtain a better result in T2D prediction compared to other binning methods.
- The results reveal high performances of state-of-the-art in deep learning algorithms, the Convolutional neural network, compared to traditional neural networks such as Multi-Layer Perceptron. Convolutional Neural networks can work efficiently even on one-dimensional data.
- Most cases, machine learning outperforms deep learning algorithms. For numeric data formed in 1D, classical machine learning reveals a robust prediction ability.
- Previous studies have not investigated the efficiency of classic machine learning with binning approaches. Our study proves by using Random Forest that it is possible be the best choice to select machine learning combining approaches to improve prediction performance on numeric species abundance datasets.

The remaining of this study, we present a short description of two considered T2D datasets in Section III. Furthermore, methods which we choose will be introduced in Section IV. Experimental Results of our proposed methods in this paper are illustrated in Section V. Finally, Section VI and Section VII discuss the results and summarize important remarks for this research.

III. Data Benchmarks for Metagenomic Analysis

We run the experiments on metagenomic abundance data that indicates how present (or absent) is an OTU (Operational taxonomic unit) in human gut. The abundance datasets are obtained using default parameters of MetaPhlAn2 described as detailed in [14].

A little more detail of the process of generating abundance shown in Fig. 1, the stool sample collected from human is fetched into machines to extract total Deoxyribo Nucleic acid (DNA). DNA then is sequenced to create millions of reads. The new generation sequencing techniques can process millions of sequencing reads in parallel. These reads are mapped to a catalog of references including all known gut microbial genes and known bacterial at levels of species, genus and so on. The techniques also indicate the presence and abundance of each gene and each species in any samples. As revealed in numerous studies, species abundance and genes abundance can distinguish patients and healthy controls. Moreover, genes and species can be leveraged to develop robust tools for diagnosis and prognosis.

We evaluated our approach on the disease of Type 2 Diabetes with two datasets. The first one (T2D1) includes 344 Chinese individuals [22], and 96 western women are in other

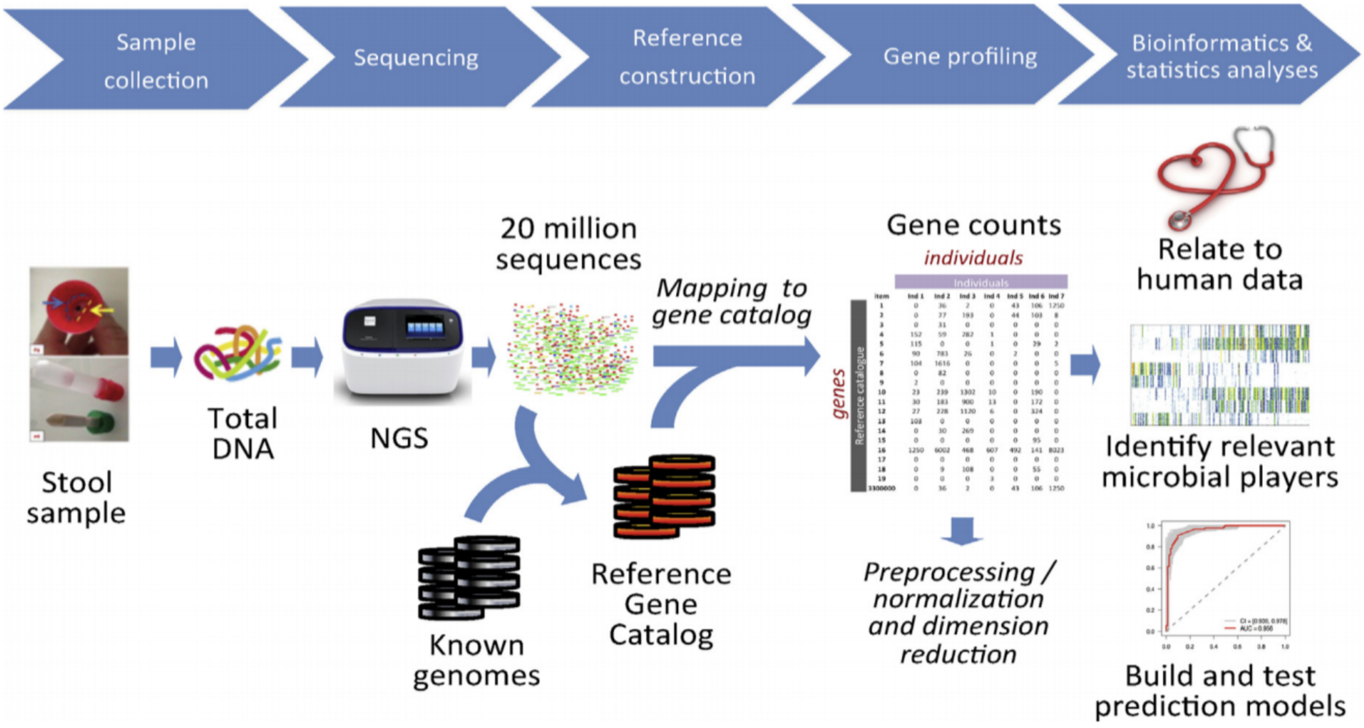


Fig. 1. Quantitative metagenomic data to explore human gut microbiome [21]

TABLE I. Binning approaches Performance Comparison in Average of Accuracy (val_acc) and Matthews correlation coefficient (val_mcc) on test sets using Multi-Layer Perceptron

Datasets	T2D1	T2D2
#Samples	344	96
#Features	572	381
#patients who affected by T2D	170	174
#controls/healthy individuals	53	43

dataset (T2D2) [23]. The datasets are characterized by bacterial species abundance. For each sample in each dataset, species abundance is a relative proportion and formed as a real number. The total abundance of all features in each sample is equal to 1. More details are shown in Table I. We consider to investigate on T2D because it is considered as one of the most changeling disease prediction tasks.

Let D be the set of considered datasets, D = $\{d_1, d_2\}$, with d_1 = T2D1, d_2 = T2D2, d = 1..2

$S_i = \{s_1, s_2, ..., s_n\}$ includes n samples corresponding to d_i

$F_i = \{f_1, f_2, ..., f_m\}$ includes m features corresponding to d_i

$P_i = \{p_1, p_2, ..., p_k\}$ includes k patients who affected by T2D corresponds to d_i

$C_i = \{c_1, c_2, ..., c_k\}$ includes x controls / healthy individuals that correspond to d_i

$$Matrix(C) = \begin{pmatrix} d_1 & S_1 & F_1 & P_1 & C_1 \\ d_2 & S_2 & F_2 & P_2 & C_2 \end{pmatrix}$$

$$= \begin{pmatrix} T2D1 & 344 & 572 & 170 & 53 \\ T2D2 & 96 & 381 & 174 & 43 \end{pmatrix}$$

Total abundance of all features in one sample is sum up to 1:

$$\sum_{i=1}^{k} f_i = 1$$

With:

- k is the number of features for a sample.
- f_i is the value of the i-th feature.

IV. Binning Approaches

A. Binning Approaches for Metagenomic Data

Some binning approaches were introduced in [24] including Species bins (SPB) based on species abundance distribution on 6 datasets, binning based on equal width and the method based on equal frequency.

- Species Bins (**SPB**) are conducted from data distribution of six metagenomic bacterial species abundance datasets related to various diseases. Authors in [25] observed that original species abundance almost follows the zero-inflated distribution. When they convert data with a scaler using log-transformed (with logarithm base 4), the scaled data is more normally-distributed (see a example of the raw species abundance and log-transformed (with logarithm base 4) of two considered datasets of T2D shown in Fig. 2).

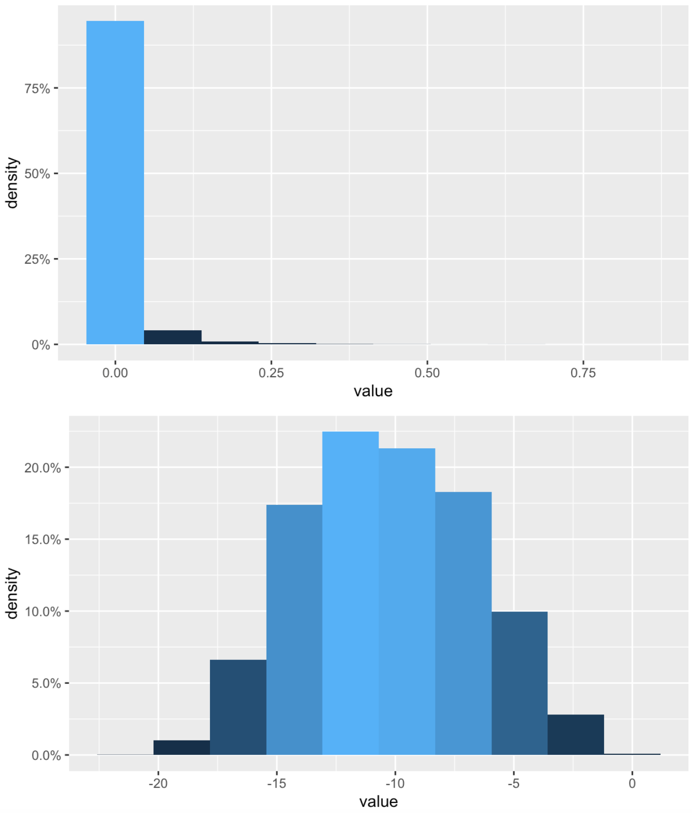


Fig. 2. Species abundance distribution of two considered T2D datasets. The top chart show original species abundance data distribution illustrates zero-inflated distribution. The other reveals a normally-distributed when we do log-transformed (with logarithm base 4) on this data.

From that, authors proposed breaks for binning where each break is the one that in the logarithm base 4 is equivalent to a fold increase from the previous bin. A little more detail, the first breaks will start at 0 and 10^{-7} (the minimum values of six considered datasets), the next break will be $4 * 10^{-7}$ and so on. This bins seem to be efficient for the prediction.

- A commonly-wided way is equal width binning (**EQW**). This technique is rather simple. The breaks are identified based on the width of the considered range of values. Let's say, we want to discretize 5 bins for a range of [Min,Max] with Min=0 and Max=0.5. The width of each bin is equal and computed by $\frac{Max-Min}{5} = 0.1$. Breaks in this example will be $0, 0.1, 0.2, 0.3, 0.4$.
- Binning based on frequency of values is also an effective method. The method is equal frequency binning (**EQF**) where each bin can contain approximately the number of elements. Therefore, the interval width can be very different. The breaks can be $0.1, 0.11, 0.2, 0.5$ and so on, for example, depending on the value distribution.
- The last binning described in this section is binary bins. This method only considers whether the value of that feature is greater 0 or not. Since it determines the Presence of feature in the samples, we also call it "**PR**".

B. *Binning based on K-means Algorithm*

With different distributions of data, the clustering algorithm is a crucial tool to identify groups in data. Determining groups for binning, we hope to improve the performance by identifying various areas which have high data density. K-means clustering is a common method in cluster analysis and data mining. The purpose of this method is to partition n elements into clusters such that each element of the cluster has the closest mean value, acting as the cluster's prototype. This method is performed based on the smallest Euclidean distance between the elements and the central element of the group. Assume each object has m attributes. Each object's properties are like coordinates of an m-dimensional space; each object is a point on that space. Euclidean distance is calculated by the formula:

$$\partial_{ji} = \sqrt{\sum_{s=1}^{m}(x_{is} - x_{js})^2}$$

With

- ai = (xi1, xi2, ... xim) i = 1..n - the ith object to be classified
- cj = (xj1, xj2, ... xjm) j = 1..k - central element group j

The central element is determined by the average of the elements in the group. Initially, these elements will be randomly selected and after each addition of objects to groups, the central elements will be recalculated. To calculate cij - the j coordinate of the group i central element, we have the formula:

$$c_{ij} = \frac{\sum_{s=1}^{t} x_{sj}}{t}$$

With:

- $j = 1..m$ (m is the number of properties)
- xsj - jth attribute of element s (s = 1..t)

Binning with K-means clustering, we will get better results than the methods mentioned earlier. Suppose we need to binning with n = 10 (the numbers of bins). This method is performed as follows:

Algorithm 1 Algorithm for identifying the list of binning breaks based on clustering algorithm, K-Means

```
Input: n - number of clusters, matrix C to find bin breaks
Output: B - array containing list of n bin breaks found
Begin
    Step 1: Initialize data
      - Convert matrix C to 1-dimensional array.
      - Remove 0 or uncountable values in array.
      - Sort the array in ascending values.
    Step 2: Using the K-means algorithm with a total
  number of clusters n - 1. We have array A containing the
  grouped elements.
    Step 3: Construct array B containing n bin breaks
      - Find n - 2 bin breaks by calculating the average of
  two boundaries in two adjacent groups.
          $B[i] = \frac{(max(A[i-1]) + min(A[i]))}{2}$
      With: $i = 1..n-1$
      - Add 0 and 1 to array B.
      - Sort the array in ascending values
End
```

For easier comparisons, all binning approaches in this study are implemented with the same number of bin (10 bins) for all classifiers. We underline that the breaks for binning are conducted using the training sets to avoid overfitting issues.

V. Experiments

For comparing the efficiency binning approaches in improving T2D prediction performance on various learning algorithms, each learning architecture is presented in each separated table. Table II gives results using MLP while Table III illustrates the performance of CNN1d. The last table (Table IV), we present the best results with Random Forest and also compare to state-of-the-art in MetAML [14]. The datasets used was described in Section III. The details of models used in the experiments and results are presented as following.

A. Learning Models for Comparison

In order to evaluate and compare the efficiency on a wide range of learning models, we propose to use 3 different learning algorithms. A state-of-the-art in machine learning is Random Forest that is implemented to run the experiments on the datasets. Moreover, as a traditional neural network, Multi-Layer Perceptron (MLP) is also leveraged for the comparison. We also evaluate one-dimensionality convolutional neural network (CNN1D) on considered datasets.

- Previous studies, most successful methods applied to numeric omics datasets are known mainly Random Forest (RF). Authors in [14] introduced MetAML using Random Forest and obtained the best results among considered algorithms. Applying the same parameters proposed in [14], we use 500 trees for this algorithm for the learning.
- The MLP is used in this study with parameters proposed in [16] including one hidden layer and 128 neural.
- CNN1D consists of one one-dimensional convolutional layer of 128 filters followed by a max pooling of 2 and ending by a fully connected layer. MLP and CNN1D use Adam optimizer function with a batch size of 16. Other parameters are also the same with a default learning rate of 0.001 and epoch patience of 5 for early stopping technique (for reducing overfitting issues).

B. Metrics for Comparison

The performances are assessed by 10-fold cross validation. We compute Average Accuracy and Average Matthews Correlation Coefficient (MCC) as performance measurement for evaluating the generalization of the classifiers. Training and test sets are exactly the same for each classifier, or we can say that the same folds are used for all classifiers. With this technique, the changes when comparing performance of any two classifiers could be computed directly as the difference in metrics within each test fold.

Accuracy is a common measurement for models's performance while MCC is considered as a good performance evaluation score for biology datasets and helps to evaluate whether the model is going well or not. As in [28], the authors said that "among the common performance evaluation scores, MCC is the only one which correctly takes into account the ratio of the confusion matrix size". Matthews correlation coefficient score is computed as following formula:

With:

- TP stands for True Positive
- TN is True Negative
- FP: False Positive
- FN: False Negative

Matthews Correlation Coefficient score is computed by:

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP+FP).(TP+FN).(TN+FP).(TN+FN)}}$$

And $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

The model reaches the best when $mcc = 1$ while the worst value is $mcc = -1$. Authors in [28] recommended using this metric for evaluating the algorithm performance.

C. Experimental Results

1) Evaluation binning approaches with MLP: We are considering two diseases T2D1 and T2D2 with results using MLP in Table II. As a result, the binning approach with K-means in both diseases achieved val_acc and val_mcc values higher than all other approaches EQW, PR, SPB. Considering dataset dataset T2D1, K-means is significantly higher than SPB. Specifically, val_acc is higher than val_acc of SPB is 0.034 and of val_mcc is 0.044. For approaches like EQW, PR or EQF, the K-Means approach returns values with relatively good disparities. Considering dataset dataset T2D2, val_acc of K-means is more than 0.069, val_mcc is 1.46 times higher than

TABLE II. Binning approaches Performance Comparison in Average of Accuracy (val_acc) and Matthews correlation coefficient (val_mcc) on test sets using Multi-Layer Perceptron

val_acc	val_mcc	Dataset	Approach
0.686	**0.379**	**T2D1**	**k-means**
0.681	0.371	T2D1	EQW
0.663	0.353	T2D1	PR
0.658	0.34	T2D1	EQF
0.652	0.335	T2D1	SPB
0.727	**0.459**	**T2D2**	**k-means**
0.714	0.437	T2D2	EQW
0.667	0.339	T2D2	PR
0.705	0.414	T2D2	SPB
0.652	0.314	T2D2	EQF

TABLE III. Binning approaches Performance Comparison in Average of Accuracy (val_acc) and Matthews correlation coefficient (val_mcc) on test sets using CNN1D

val_acc	val_mcc	Dataset	Approach
0.692	**0.392**	**T2D1**	**k-means**
0.678	0.363	T2D1	EQW
0.677	0.367	T2D1	PR
0.652	0.323	T2D1	EQF
0.649	0.316	T2D1	SPB
0.740	**0.473**	**T2D2**	**k-means**
0.707	0.413	T2D2	EQW
0.700	0.397	T2D2	PR
0.687	0.382	T2D2	SPB
0.674	0.346	T2D2	EQF

EQF. The value of EQW in this disease is the second most in approach and is 0.022 different from when using K-Means. In summary, the results when binning with K-Means cluster using Multi-Layer Perceptron, we will get the best results compared to the remaining methods.

2) *Evaluation binning approaches with Convolutional Neural Network on 1D data:* Table III shows the performance using CNN1D. When using the One-Dimensional Convolutional Neural Network, the results of K-Means are 0.692 for val_acc, 0.740 for val_mcc, respectively. Both results are better than using Multi-Layer Perceptron (val_acc = 0.686, val_mcc = 0.727). In T2D1, the result of K-Means is much higher than the next EQW value, namely 0.014 difference for val_acc and 0.076 for val_mcc compared to K-Means. The value of val_acc of K-Means compared to the lowest value in this disease of SPB is 0.076 and of val_mcc is 0.043. In T2D2, the lowest valued approach for this disease is EQF. Val_acc value is more than 0.066, val_mcc of K-Means is 1.367 more than EQF. The difference between the values of EQW and K-Means is quite good, respectively 0.033 for val_acc, 0.06 for val_mcc. In summary, when using the One-Dimensional Convolutional Neural Network, the K-Means approach results in better results when using the Multi-Layer Perceptron and this result is still the best result compared to the other approach.

3) *Random Forest obtains promising results with the proposed binning, compared to state-of-the-art MetAML:* We also used the Random Forest for results comparison in Table III. Similar to the previous two tables, when binning with K-means we obtain very good results compared to using other approaches. A previously used framework, MetAML, K-means, gave val_acc more than 0.036 for T2D1 and 0.056 for T2D2. Considering T2D1, K-means val_acc is more than 0.04 and val_mcc is 0.07 more than SPB. The second result in the table for both diseases is the PR approach. The difference in value between K-means and PR is quite good. K-means has val_acc more than 0.014, val_mcc is more than 0.017 than PR. Considering T2D2, val_acc is 0.107 and val_mcc is 1.683 times higher than SPB results. K-means has val_acc more than 0.023, val_mcc is more than 0.032 than PR. In short, when choosing K-means as an approach, we will get better results than some common approaches such as PR, EQW, EQF or SPB, especially the approach used was MetAML.

TABLE IV. Binning approaches Performance Comparison in Average of Accuracy (val_acc) and Matthews correlation coefficient (val_mcc) on test sets using Random Forest

val_acc	val_mcc	Dataset	Approach
0.700	**0.400**	**T2D1**	**k-means**
0.686	0.383	T2D1	PR
0.680	0.370	T2D1	EQF
0.674	0.357	T2D1	EQW
0.660	0.330	T2D1	SPB
0.664		T2D1	*MetAML*
0.759	**0.515**	**T2D2**	**k-means**
0.736	0.483	T2D2	PR
0.720	0.440	T2D2	EQW
0.690	0.370	T2D2	EQF
0.652	0.306	T2D2	SPB
0.703			*MetAML*

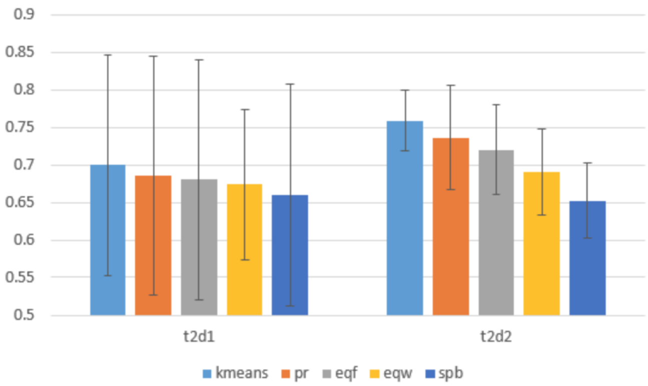


Fig. 3. Performance Comparison in Average Accuracy of different binning approaches including EQF, EQW, K-means, PR and SPB. Standard deviations are shown in error bar.

4) *Random Forest obtains better results compared to neural networks:* The chart in Fig. 3 shows the results being conducted from two datasets of T2D. We use five approaches for testing, namely, EQF, EQW, K-Means, PR, SPB. Considering T2D1 disease, the K-means approach has the largest Average Accuracy value, reaching 0.7. SPB has a value of Average Accuracy is 0.66, this is the smallest value and smaller than K-Means 0.34. Similarly, for T2D2 disease, the Average Accuracy of K-Means value is 0.759, the highest among the remaining approaches. This value is higher than the next PR value of 0.023. The Average Accuracy of SPB is less than 0.107 compared to K-Means.

The chart in Fig. 4 shows the results Average MCC value on 2 datasets of T2D and 5 approaches. K-Means has the highest Average MCC value on both datasets and 0.4 for T2D1 and 0.515 for T2D2. Average MCC value of K-Means greater than SPB in T2D1 is 0.07, 1,683 times that of T2D2. The disparity with the next high value of PR is also quite clear,

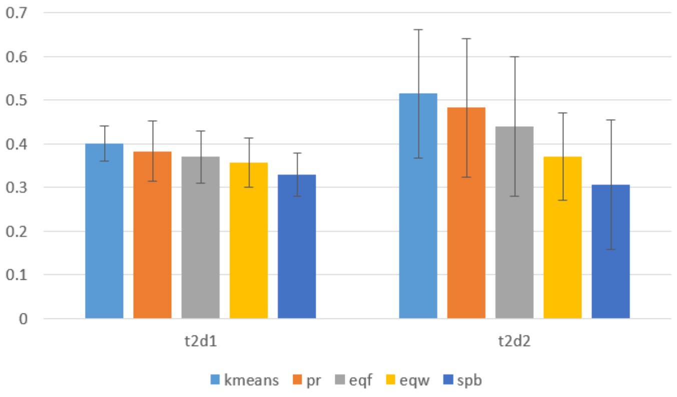


Fig. 4. Performance Comparison in Average MCC of different binning approaches including EQF, EQW, K-means, PR and SPB.

namely, 0.017 for T2D1 and 0.032 for T2D2.

VI. Discussion

From collected results, we can see that RF obtains the best among considered models. These results are similar to [25] where authors also have attempted to apply deep learning but the performance in T2D disease is still worse than RF. This reflects a fact as mentioned in [26]: “the deep learning approaches may not be suitable for metagenomic applications”. As stated in [27], we are facing challenges when applying deep learning to solve biological and clinical tasks because of limited data availability, result interpretation and hyper-parameters tuning for deep learning algorithms.

Although PR only considers whether a bacterial species exists in a patient or, it revels a better performance (using RF) than several other binning methods such as SPB, EQW, EQF. From results, we can propose medical examinations for T2D only determining the existence of bacterial species in human body for the diagnosis. These examinations can be simpler than computing quantitative compositions of bacterial.

In most situations, SPB performs poor performance compared to the others because SPB was conducted from species abundance distribution from various diseases. Each disease should be considered independently because one disease can have its own complexity, characteristics as well as data density.

VII. Conclusion

We introduce a novel binning approach using a classical clustering algorithm such as K-means. As shown from the comparison results among considered existing binning approaches such as binning based on species distribution, based on width and frequency and binary bins, we can see the encouraging results in use of clustering methods for identifying breaks for binning to enhance the prediction performance.

The analysis of two architectures of one-dimensional convolutional neural network and Multi-layer Perceptron shows that convolutional neural network not only achieve a good performance on images but also obtain a promising result compared to traditional neural network such as MLP.

As some results in previous studies, classic machine learning such as Random Forest still works better more complex models such as MLP and CNN1D in T2D diagnosis by metagenomic data. Further research can investigate more deeper and sophisticated models to improve the performance.

Using classic clustering algorithm K-means with default parameters in binning gives encouraging results. This could promote studies to go deeper in use of clustering methods to generate breaks for binning. This illustrate that there are potentials in exploring density data to improve not only for T2D disease but also for other diseases.

References

[1] Kevin Chen, Lior Pachter. Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities. 2005.

[2] DeLong EF Microbial population genomics and ecology. Curr Opin Microbiol 5: 520–524. 2002.

[3] Handelsman J, Metagenomics: Application of genomics to uncultured microorganisms. Microbiol Mo lBiol Rev 68: 669-684. 2004

[4] Riesenfeld CS, Schloss P, Handelsman J, Metagenomics: Genomic analysis of microbial communities. Annu Rev Genet 38: 525–552. 2004.

[5] Rodriguez-Valera F, Environmental genomics, the big picture? FEMS Microbiol Lett 231: 153–158. 2004.

[6] Streit WR, Schmitz RA, Metagenomics—The key to the uncultured microbes. Curr Opin Microbiol 7: 492–498. 2004.

[7] Maja Fabijanić and Kristian Vlahoviček, Oliviero Carugo and Frank Eisenhaber (eds.), Data Mining Techniques for the Life Sciences, Methods in Molecular Biology, vol. 1415, DOI 10.1007/978-1-4939-3572-7_26, © Springer Science+Business Media New York 2016.

[8] Edwards RA, Rohwer F, Viral metagenomics. Nat Rev Microbiol 3: 504–510. 2005.

[9] NIH HMP Working Group, Peterson J, Garges S et al, The NIH Human Microbiome Project. Genome Res 19:2317– 2323. 2009. doi:10.1101/gr.096651.109. 2009.

[10] Garrett WS, Gallini CA, Yatsunenko T et al, Enterobacteriaceae act in concert with the gut microbiota to induce spontaneous and maternally transmitted colitis. Cell Host Microbe 8:292–300. doi:10.1016/j.chom.2010.08.004. 2010.

[11] Karlsson FH, Fåk F, Nookaew I et al, Symptomatic atherosclerosis is associated with an altered gut metagenome. Nat Commun 3:1245. doi:10.1038/ncomms2266. 2012.

[12] Qin N, Yang F, Li A et al, Alterations of the human gut microbiome in liver cirrhosis. Nature 513:59–64. doi:10.1038/nature13568. 2014.

[13] Turnbaugh PJ, Gordon JI, The core gut microbiome, energy balance and obesity. J Physiol 587:4153–4158. doi:10.1113/ jphysiol.2009.174136. 2009.

[14] E. Pasolli, D. T. Truong, F. Malik, L. Waldron & N. Segata; Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights; PLoS Comput. Biol. 12, p. e1004977. 2016.

[15] Steve Miller, Charles Chiu, Kyle G. Rodino, Melissa B. Miller; Point-Counterpoint: Should We Be Performing Metagenomic Next-Generation Sequencing for Infectious Disease Diagnosis in the Clinical Laboratory?. DOI: 10.1128/JCM.01739-19. Journal of Clinical Microbiology. 2020.

[16] Thanh Hai Nguyen, Jean-Daniel Zucker. Enhancing Metagenome-based Disease Prediction by Unsupervised Binning Approaches. The 2019 11th International Conference on Knowledge and Systems Engineering (KSE-IEEE), ISBN: 978-1-7281-3003-3, pp 381-385. 2019.

[17] Hanaa M. Hussain et al. FPGA implementation of K-means algorithm for bioinformatics application: An accelerated approach to clustering Microarray data. 2011 NASA/ESA Conference on Adaptive Hardware and Systems (AHS). 2011.

[18] Timothy et al. K-Means Clustering of Biological Sequences. ADCS 2017: Proceedings of the 22nd Australasian Document Computing Symposium. 2017.

[19] Jasmin T. Jose1 et al. Case Study on Enhanced K-Means Algorithm for Bioinformatics Data Clustering. International Journal of Applied Engineering Research ISSN 0973-4562. 2017.

[20] Vijayalakshmi K., Padmavathamma M. (2019) Design and Implementation of Modified Sparse K-Means Clustering Method for Gene Selection of T2DM. In: Computational Intelligence and Big Data Analytics. SpringerBriefs in Applied Sciences and Technology. Springer, Singapore. 2019.

[21] Stanislav Dusko Ehrlich. The human gut microbiome impacts health and disease. PubMed. 339(7-8):319-23. doi: 10.1016/j.crvi.2016.04.008. PMID: 27236827. 2016

[22] Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature 2013;498(7452):99–103. pmid:23719380. 2013.

[23] Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A. 2013 metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 2012;490(7418):55–60. pmid:23023125. 2012.

[24] Le Chatelier E, Nielsen T, Qin J et al Richness of human gut microbiome correlates with metabolic markers. Nature 500:541–546. doi:10.1038/nature12506. 2013.

[25] Thanh Hai Nguyen et al.; Disease Classification in Metagenomics with 2D Embeddings and Deep Learning; In Proceedings of CAp, France 2018.

[26] G. Ditzler, R. Polikar & G. Rosen; Multi-Layer and Recursive Neural Net- works for Metagenomic Classification; IEEE Trans. Nanobioscience 114, p. 608–616. 2015.

[27] Fioravanti, D., Giarratano, Y., Maggio, V. et al. Phylogenetic convolutional neural networks in metagenomics. BMC Bioinformatics 19, 49. https://doi.org/10.1186/s12859-018-2033-5. 2018.

[28] Baghban, H. and Rahmani, A.M. A heuristic on job scheduling in grid computing environment. In Grid and Cooperative Computing, 2008. GCC'08. Seventh International Conference on (pp. 141-146). IEEE. October, 2008.

Enhanced Performance of the Automatic Learning Style Detection Model using a Combination of Modified K-Means Algorithm and Naive Bayesian

Nurul Hidayat[1]
Department of Computer Science and Electronics
Universitas Gadjah Mada, Yogyakarta, Indonesia
Department of Informatics
Jenderal Soedirman University, Purwokerto, Indonesia

Retantyo Wardoyo[2§]
Department of Computer Science and Electronics
Universitas Gadjah Mada, Yogyakarta, Indonesia

Azhari SN[3]
Department of Computer Science and Electronics
Universitas Gadjah Mada, Yogyakarta, Indonesia

Herman Dwi Surjono[4]
Department of Electronics Engineering Education
Yogyakarta State University(UNY), Indonesia

Abstract—**Learning Management System (LMS) is well designed and operated by an exceptional teaching team, but LMS does not consider the needs and characteristics of each student's learning style. The LMS has not yet provided a feature to detect student diversity, but LMS has a track record of student learning activities known as log files. This study proposes a detection model of student's learning styles by utilizing information on log file data consisting of four processes. The first process is pre-processing to get 29 features that are used as the input in the clustering process. The second process is clustering using a modified K-Means algorithm to get a label from each test data set before the classification process is carried out. The third process is detecting learning styles from each data set using the Naive Bayesian classification algorithm, and finally, the analysis of the performance of the proposed model. The test results using the validity value of the Davies-Bouldin Index (DBI) matrix indicate that the modified K-Means algorithm achieved 2.54 DBI, higher than that of original K-Means with 2.39 DBI. Besides having high validity, it also makes the algorithm more stable than the original K-Means algorithm because the labels of each dataset do not change. The improved performance of the clustering algorithm also increases the values of precision, recall, and accuracy of the automatic learning style detection model proposed in this study. The average precision value rises from 65.42% to 71.09%, the value of recall increases from 72.09% to 80.23%, and the value of accuracy increases from 67.06% to 71.60%.**

Keywords*—*Learning Management System; log file, K-Means; Davies-Bouldin Index

I. Introduction

The rapidly developing information and communication technology currently offer excellent potential to overcome the problem of equitable access to quality learning in Higher Education through the Learning Management System (LMS). LMS is a software application or web-based technology used to plan, implement, and assess a particular learning process. Although LMS is well designed and operated by an exceptional teaching team, the learning process through the LMS has a weakness: an inability to personalize the learning [1]. This is caused by the nature of LMS that provides the same content for all students in a given course. Each student has a different learning style and can learn better in different ways. Different geographical and socio-cultural locations of students will certainly form different learning styles [2]. Learning styles can influence and motivate students to take lessons. One of the main things that needs to be considered in learning with e-learning systems is individual learning styles that vary in LMS. For example, the contents of my course, subjects, and student behavior and online learning experiences can influence learning styles.

Currently, several learning styles have been used, such as Honey and Mumford, Kolbs, Felder Silverman Learning Style Model (FSLSM), and VAK [3]. There are also Gregorc's learning styles, Riding cognitive styles, and Myer-Briggs Type Indicator [4]. FSLSM is the most widely used learning style in the education system, which shows a high level of reliability, internal consistency, and validity [4]–[8]. This model defines student learning styles into four different dimensions (Active/Reflective, Sensitive/Intuitive, Visual/Verbal, Sequential/Global) based on student behavior patterns that use E-learning systems [6]. Students with a strong preference for a particular learning style may have learning difficulties if the teaching style does not match the student's learning style. To reach the goal of equal education successfully, the development of LMS is needed so that they present learning sources with the context and the learning process that is suitable for the student's learning styles to improve their performance. Therefore, it needs a way to classify the learning styles of each student by detecting their learning styles.

Currently, the detection of learning styles can be divided into two approaches, namely, static and automatic [3]. Static approach is a learning style detection approach that is done by using a questionnaire [9]–[13]. Students need to fill out questionnaires to identify their learning style. Constraints faced by students in the process of filling out this questionnaire can take a long time and tend to students only have a target

[§] Corresponding author's Email ID: rw@ugm.ac.id of completing questions without understanding the purpose of filling out the questionnaire [14]. Students often answer questionnaires with unresponsiveness, so the results of detection using this questionnaire tend to be inaccurate [15]. Therefore, the focus of research using a static approach lies in the reliability and validity of the Index of Learning Style (ILS) instruments.

The second approach is automatic, which is based on actual behavior patterns during online learning. The approach is based on personality factors, behavioral factors and time. The automatic detection process is far more accurate, dynamic, and comprehensive than static detection because the interaction process is directly recorded without being noticed by the participant by using log file data and does not require special time [3]. Two methods can be used to determine learning styles automatically, namely: data-driven methods and literature-based methods [16]. Data-driven methods aim to build classification models that copy ILS instruments and use sample data to build models. Some classification techniques that are widely used to detect this learning style include Neural Network [17]–[19], Decision Tree [1], [6], [7], [20], [21], and Bayesian Network [22]–[26]. The literature-based method uses student behavior and actions with the system to identify their learning preferences. Some studies that use a literature-based approach include [2], [15], [17], [27], [28]. The developed method uses simple rule-based methods to calculate learning styles from the number of suitable instructions and does not involve system design. This approach still has problems in estimating the importance of various instructions used to calculate learning style preferences. Also, it requires some knowledge of psychology and cognitive science to estimate the importance of calculating learning style preferences.

Based on reviews [3] the most widely used automated approach model is the Bayesian Network model. Bayesian networks can naturally represent probabilistic information, efficiency, and support to encode uncertain expert knowledge. Also, Bayesian Network makes it possible to model quantitative and qualitative information about student behavior [22]. According to [29], in general, the Bayesian Network is too complicated for small data sets and is easy to be overfitted. This problem can be avoided by using the Naive Bayesian (NB) algorithm. The advantages of the classification using the NB algorithm are that it is easy to build because the structure is given a priority, and there are no learning procedures, as well as an efficient classification process. Both of these advantages are obtained by assuming that all features are independent of one to another. However, the requirements of each node must be separate, making the NB structure produce low accuracy. One of the improvements in the accuracy of the NB structure is to determine the appropriate class label before classification. One method that can be used to get class labels is the clustering method.

Clustering is very suitable for grouping data, which class labels are difficult to obtain at the time of feature generation. Many clustering algorithms are used to get class labels. One of the most used clustering algorithms is the K-Means algorithm. This is because the K-Means algorithm is easy to implement, the time needed to carry out this learning is relatively fast, easy to adapt, and is very suitable for clustering with a large number of groups. However, the K-Means Algorithm also has weaknesses, namely, the results of clustering are less than optimal due to the initial centroid in the initialization process are chosen randomly. If implemented with quite a lot of features, then the K-Means algorithm also has a problem known as the curse of dimensionality [30]. Therefore, to improve the performance of the K-Means algorithm, it can be developed by enhancing the initial centroid selection process.

According to [25], most studies detecting FSLSM learning styles group learning styles into eight combinations of learning styles. If observed from the FSLSM learning style model consisting of 4 dimensions with each dimension having two categories, then it is possible to have 16 combinations of learning styles. Therefore, in this study, a modification of the proposed K-Means algorithm was used to classify FSLSM learning style models to 16 groups before learning styles were detected using classification methods.

In this paper, the proposed improvement of the FSLSM learning style detection model is carried out by combining the modification of the K-Means algorithm with the Naive Bayesian classification algorithm. The detection process of the proposed learning style model consists of four methods, namely pre-processing, which aims to translate the data log file to several characteristics such as skills, level of knowledge, preferences, and learning styles that are considered to affect the learning process of students directly. This process produces in 29 features used for the grouping process of the dataset derived from the participants of the Education for Professional Teachers held by the Ministry of Research and Technology for teachers of English subject with 500 data. The second process is grouping using a modified K-Means algorithm to obtain cluster labels from each test data set. The fourth process is to detect learning styles from each data set using the Naive Bayesian classification algorithm, and finally to analyze the performance of the proposed automatic learning style detection model.

This paper is organized as follows. Section 2 discusses related work. Section 3 elaborates the proposed model, followed by Section 4 containing analysis of performance evaluation of the proposed modified K-Means algorithm. Finally, Section 5 concludes this paper.

II. Related Works

Conventional learning generally uses a one-to-many tutor approach, where lecturers deliver material without looking at the diversity of students' knowledge, so the content offered is not optimal. One solution that can be used is to use a one-to-one tutor approach, but the method can be said to be impossible to be applied to conventional learning because of time constraints. The development of information technology has an impact on education, namely the use of a Learning Management System (LMS). The emergence of LMS has the potential to be applied to a one-to-one tutor approach because LMS provides easy access by lecturers and students without being bounded by time.

Learning style is an essential factor that plays a role in individual student's learning in any learning environment. Each student has a different learning style and different ways to understand, process, maintain, and understand new information. Learning style is a way for students to follow learning effectively and efficiently. Currently, there are 4 models of learning styles that are most widely used, namely: Honey and Mumford, Kolbs, VAK, and Felder-Silverman Learning Style Model (FSLSM).

Honey and Mumford's learning style model introduces the concept of learning style based on the description of attitudes and behaviors that determines the way of learning preferred by learners using the Learning Style Questionnaire (LSQ) [31]. LSQ is designed to investigate the relative strengths of four different learning style dimensions from Honey and Mumford [32], namely, Activity, Reflector, Theory, and Pragmatic. Research carried out to determine Honey and Mumford's learning style models focuses on learning models. Research conducted by [31] produced a valid and reliable research questionnaire. Likewise, research conducted by [32] states that the research is significant following the principles of learning styles proposed by Honey and Mumford, which are statistically tested.

Kolbs, the learning style model, introduces the Learning Styles Inventory to identify individual learning styles [33]. Learning Styles Inventory is understood as a four-dimensional cycle consisting of Concrete experience (CE), reflective observation (RO), Abstract Conceptualization (AC), and Active Experimentation (AE). Research on Kolbs' learning style focuses on behavior by using the concept of Questionnaire [34], [35], and log file [36]. Research conducted by [34] aims to detect learning styles using Kolb's 4-dimensional and 9-dimensional, while [35] to detect learner's learning styles in LMS automatically uses the Naive Bayesian technique to replace the Kolbs' Learning Styles Inventory (KLSI). Research conducted by [36] aims to classify learner's learning styles based on the Decision Tree algorithm using the log data file.

The VAK learning style model categorizes learners' learning styles based on three dimensions [37], namely: Visual, Auditory, and Kinesthetic. VAK learning style research is mostly aimed at Behavior using Literature Base, Questionnaire, and Latent Semantic Indexing. VAK architecture to detects learning styles based on student behavior using simple rule-based techniques introduced by [37]. The research aimed at identifying VAK learning styles were carried out by [38] using the Decision Tree C4.5 algorithm on questionnaire data. Meanwhile, [39] predicted VAK learning styles using the artificial neural network (ANN) method is Latent Semantic Indexing.

The Felder-Silverman Learning Style Model (FSLSM) uses the notion of dimensions where each dimension contains two opposing categories, and each student has a dominant preference in each category of dimension. The four dimensions of the FSLSM are Processing (Active/Reflective), Perception (Sensing/Intuitive), Input (Visual/Verbal), and Understanding (Sequential/Global). FSLSM allows Learning Style (LS) to be measured based on the Index of Learning Style (ILS). Therefore, by using ILS, we can link the LS to the appropriate learning objects. The FSLSM learning style research model is mostly about behavior using log file data using different classification algorithms. Research by [40] classifies FSLSM learning styles using Fuzzy Cognitive Maps (FCMs), while [6] uses Decision Tree.

Some researchers also focus on finding appropriate learning style models in LMS, including [33], by comparing three models of Honey and Mumford's learning style questionnaire, Kolb, and FSLSM. The test results are measured based on how easy the questions to be understood, the time needed to fill out the questionnaire, and how the results are presented. The measurement results stated 67% of respondents understood the ILS FSLSM learning style model more easily and required less time than Honey and Mumford's and Kolb's methods. Whereas [41] evaluated the adaptive E-learning system based on the VAK learning style with FSLSM that had been developed using the LMS model. Based on the explanation, most of the researchers mapped the student's learning style model to the FSLSM learning style model. Also, the results of the study [33] stated that the FSLSM questionnaire model was easier to understand and needed more time to complete the assessment. Therefore, this study uses the FSLSM learning style model to automatically detect students' learning patterns in LMS for the participants of the Education of Professional Teachers (PPG) SPADA Kemenristekdikti for teachers of English language subject.

Several learning-style models have been introduced, such as the Honey and Mumford, Kolbs, FSLSM, and VAK models, but the main problem of learning through LMS is how to identify student's learning styles that fit the model. The issue of learning style can be solved using two main approaches, namely, the static and automatic approaches [42]. Learning style detection research using a static method is mostly used to measure the reliability and validity of the Index of Learning Style (ILS) instruments [11]–[13]. The results of the study to detect learning styles using a static approach show the value of preference in each low dimension, i.e., the average of each dimension is below 50% [10], [43]. This shows some limitations of the static approach; the first is related to the lack of student's motivation to fill out questionnaires and lack of awareness of their learning preferences.

The second problem is that filling out questionnaires is very tedious and takes up student's time because there are usually quite a lot of items on the polls. The third problem is students can be influenced by the way the questionnaire is formulated, which can affect students in providing the answers [3]. Based on the weaknesses of the static learning style approach, subsequently, many researchers conducted research using an automatic method.

Research on learning style detection using an automatic approach mostly uses data-driven, which is the data log files. Besides, the study conducted aims to determine the best classification algorithm among Algorithm Decision Tree (J48), Artificial Neural Network, and Support Vector Machine to detect student's learning styles into eight learning styles FSLSM [5], [17], [18], [21], [23], [26], [44]. The results of the comparison of the performance of the classification algorithm to detect FSLSM learning styles provide the Naive Bayesian algorithm better than other Data Mining algorithms. However, the Naive Bayesian algorithm has precision and accuracy values that are still below the algorithm of Artificial Neural Network with J48. This proves that the classification approach can be very accurate, depending on the available data. Improving the accuracy of the classification model can be done by determining the appropriate class label using the clustering method. Therefore, in this study, the performance improvement of the Naive Bayesian algorithm for detecting learning styles automatically using an algorithm for grouping log file data based on the FSLSM dimensions, namely the modified K-Means algorithm before classification.

III. Proposed Method

An outline of the proposed automatic learning style detection model using the merging of K-Means logarithm modification with Naive Bayesian is shown in Fig. 1.

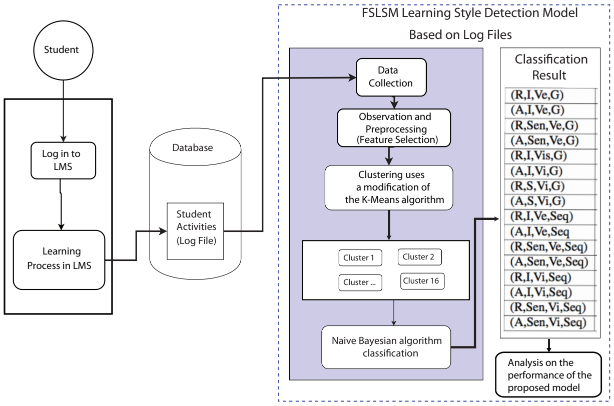


Fig. 1. Automatic learning style detection of the proposed model

Based on Fig. 1 this research process consists of four main steps, namely: observation and pre-processing, the process of grouping learning style models using modified K-Means algorithm, classification using the Naive Bayesian algorithm and testing of the proposed model.

A. Observation and Pre-processing

The purpose of this step is to get features that correlate with the type of FSLSM learning style. This stage analyzes the data log file based on four dimensions of the FSLSM model. The observation process was carried out on 47 files from the log file data to determine the features of the log data file. Logfile data is formed automatically when students use the LMS system. The system records all activities in the form of chat, forums, quizzes, exercises, assignments, examination submissions, frequency of accessing subject matter, etc. These activities then formed the features. Furthermore, each file that has features that correlate with features needed was sorted to obtain 4 dimensions of FSLSM. Based on the observations of 47 files from the log file data, 22 files are containing 423 features that may correlate with the features of FSLSM.

The pre-processing was carried out on 22 files from the log file data by removing:

- Duplicate data, thereby reducing the number of rows and columns from the data set.
- N.A. data for each user id should have recorded data on the activities of the use of the learning system. Still, the information is not widely available, so there is a lot of incomplete data.
- Removes rows of data that cannot be related to data rows in other tables because they do not share the same column.
- Determine user ID of PPG SPADA participants Kemenristekdikti teachers teaching English subjects as much as 500 data randomly.

The pre-processing process resulted 29 features consisting of 9 dimensions of processing features, 9 features of perception dimension, 6 features of the input dimension, and 5 features of understanding dimension, as shown in Table I.

TABLE I. The learning style detection feature results in preprocessing

Dimension	Feature Name	Description of Student Behavior
Processing	*Online_Forum*	F1-Post messages and reply to messages F2-Read the message F3-Never use
	E-mail	E1-Very often used E2-Sometimes E3-Never use
	Online_Chat	C1-Very often used C2-Sometimes C3-Never use
Perception	*Exam_revision*	R1-Test scores more than 75 R2-Test scores between 25-75 R3-Test score is less than 25
	Assessment	A1-Following the quiz more than 7 times A2-Following the quiz a little (2-7 times) A3-Following the quiz less than 2 times
	Exercise	Ex1-Number of exercises to follow: many (more than 7 times) Ex2-Number of exercises to take: a little (2-7 times) Ex3-Number of exercises followed: less than 2 times
Input	Input_Teks	I1-Text-based learning objects used: many (more than 75%) I2-Text based learning objects used: few (25-75%) I3-Text based learning object used: none
	Input_Multimedia	M1-Multimedia-based learning objects (audio, video, images) used: many (more than 75%) M2-Multimedia-based learning objects (audio, video, images) used: a little (25-75%) M3-Multimedia-based learning objects (audio, video, images) used: none
Understanding	*Exam_results*	Er1-Test scores: more than 75 Er2-Test scores: 25-75 Er3-Test scores: less than 25
	Long stay at LMS	L1-Average of more than 200 minutes L2-An average of less than 200 minutes

B. The Process of Clustering Learning Style Models Using Modified K-Means Algorithms

Modified K-Means algorithm is used to obtain labels from the learning style model for detection are shown in Fig. 2. Modifications of the K-Means algorithm are performed to determine the data set to be selected as the initial centroid.

The process of clustering using algorithms K-Means can be explained as follows:

1) Early initialization and centroid determination process: This step is used to determine the number of clusters (K) and the objective function value (FO). This research uses $K = 16$ according to the FSLSM learning style model grouping, as shown in Table II.

TABLE II. Combination of FSLSM Learning Styles

Cluster	Learning Style	Cluster	Learning Style
1	(A,S,Vi,Seq)	9	(R,S,Vi,Seq)
2	(A,S,Vi,G)	10	(R,S,Vi,G)
3	(A,S,Ve,Seq)	11	(R,S,Ve, Seq)
4	(A,S,Ve,G)	12	(R,S,Ve,G)
5	(A,I,Vi,Seq)	13	(R,I,Vi,Seq)
6	(A,I,Vi,G)	14	(R,I,Vi,G)
7	(A,I,Ve,Seq)	15	(R,I,Ve,Seq)
8	(A,I,Ve,G)	16	(R,I,Ve,G)

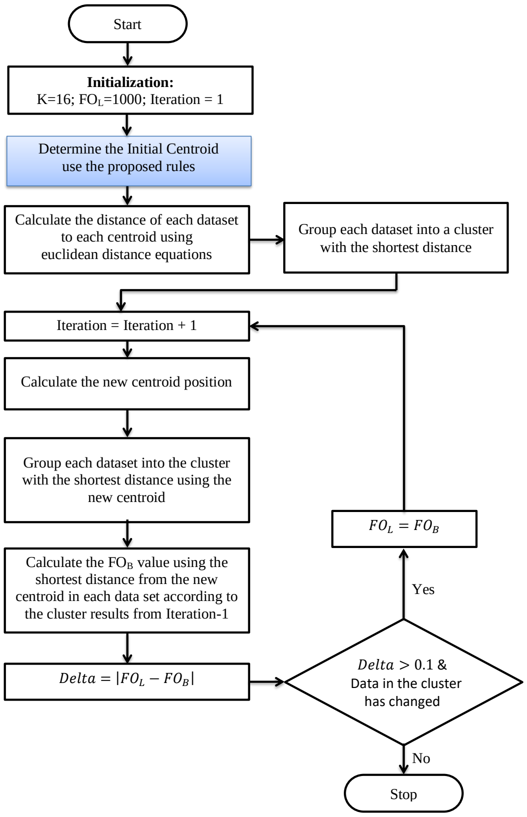

Start
Initialization:
K=16; FO$_L$=1000; Iteration = 1
Determine the Initial Centroid use the proposed rules
Calculate the distance of each dataset to each centroid using euclidean distance equations
Group each dataset into a cluster with the shortest distance
Iteration = Iteration + 1
Calculate the new centroid position
Group each dataset into the cluster with the shortest distance using the new centroid
Calculate the FO$_B$ value using the shortest distance from the new centroid in each data set according to the cluster results from Iteration-1
$Delta = |FO_L - FO_B|$
$Delta > 0.1$ & Data in the cluster has changed
Yes
$FO_L = FO_B$
No
Stop

Fig. 2. Modified K-Means algorithms

The value of FO is determined by a sufficiently high value, for example, 1000. The purpose of determining the initial value of FO is that the iteration process is not only done once so that the clustering results can be optimal.

The next step is the process of determining the initial centroid. This step is the core of the proposed modified K-Means algorithm . Modifications made are in the process of determining the initial centroid using rules established by the author. In contrast, in the original K-Means algorithm, the initial centroid determination is done by selecting K random data set.

The rules used to determine the initial centroid are 16 data sets that are carried out by identifying all data sets that meet the FSLSM learning style model criteria in Table II, which was discovered first. The criteria for each FSLSM learning style model are available in Table II, which was used to determine the initial centroid using the following rules:

- The learning style in the Processing dimension $(D1_i)$ determined by equation (1).

$$D1_i = \begin{cases} A & if\ P1_i > 3 \\ R & if\ P1_i \leq 3 \end{cases} \tag{1}$$

with i is the dataset number i, A is an Active learning style category, R is a Reflective learning style category, and $P1_i$ is the value of preference at $D1_i$ obtained from the equation (2).

$$P1_i = \frac{(FH_i + EH_i + CH_i))}{3} \tag{2}$$

with the provision of:

$$FH_i = \begin{cases} max(F(i,:)) & if\ max(F(i,:)) = F(i,1) \\ & or\ max(F(i,:)) = F(i,2) \\ 0 & if\ max(F(i,:)) = F(i,3) \end{cases}$$

$$EH_i = \begin{cases} max(E(i,:)) & if\ max(E(i,:)) = E(i,1) \\ & or\ max(E(i,:)) = E(i,2) \\ 0 & if\ max(E(i,:)) = E(i,3) \end{cases}$$

$$CH_i = \begin{cases} max(C(i,:)) & if\ max(C(i,:)) = C(i,1) \\ & or\ max(C(i,:)) = C(i,2) \\ 0 & if\ max(C(i,:)) = C(i,3) \end{cases}$$

FH, EH, and CH : is the maximum value of each preference in the Processing dimension, respectively, the Forum feature (F), E-mail feature (E), and Online_chat features (C).

- The learning style on the Perception dimension $(D2_i)$ determined based on the equation (3).

$$D2_i = \begin{cases} S & if\ P2_i > 3 \\ I & if\ P2_i \leq 3 \end{cases} \tag{3}$$

with i is the dataset number i, S: is the Sensing learning style category, I: is an Intuitive learning style, and $P2_i$: is the value of preference at $D2_i$ obtained from the equation (4).

$$P2_i = \frac{(RH_i + AH_i + ExH_i))}{3} \tag{4}$$

with the provision of:

$$RH_i = \begin{cases} max(R(i,:)) & if\ max(R(i,:)) = R(i,1) \\ & or\ max(R(i,:)) = R(i,2) \\ 0 & if\ max(R(i,:)) = R(i,3) \end{cases}$$

$$AH_i = \begin{cases} max(A(i,:)) & if\ max(A(i,:)) = A(i,1) \\ & or\ max(A(i,:)) = A(i,2) \\ 0 & if\ max(A(i,:)) = A(i,3) \end{cases}$$

$$ExH_i = \begin{cases} max(Ex(i,:)) & if\ max(Ex(i,:)) = Ex(i,1) \\ & or\ max(Ex(i,:)) = Ex(i,2) \\ 0 & if\ max(Ex(i,:)) = Ex(i,3) \end{cases}$$

RH, AH, and ExH : is the maximum value of each preference in the Perception dimension i.e., successively is the Exam revision feature (R), Assessment features (A), and Exercise features (Ex).

- The learning styles in the Input dimension $(D3_i)$ determined based on the equation (5).

$$D3_i = \begin{cases} Vi & if\ P3_i > 3 \\ Ve & if\ P3_i \leq 3 \end{cases} \tag{5}$$

with i is the dataset number i, Vi: is a category of Visual learning styles, Ve: is a Verbal learning style category, and $P3_i$: is the preference value at $D3_i$ obtained from the equation (6).

$$P3_i = \frac{(IH_i + MH_i))}{2}; \tag{6}$$

with the provision of:

$$IH_i = \begin{cases} max(I(i,:)) & if\ max(I(i,:)) = I(i,2) \\ & or\ max(I(i,:)) = I(i,3) \\ 0 & if\ max(I(i,:)) = R(i,1) \end{cases}$$

$$MH_i = \begin{cases} max(M(i,:)) & if\ max(M(i,:)) = M(i,1) \\ & or\ max(M(i,:)) = M(i,2) \\ 0 & if\ max(M(i,:)) = M(i,3) \end{cases}$$

IH and MH : is the maximum value of each preference in the Input dimension, which successively is Input_teks feature (I) and Input_Multimedia features (M).

- The learning styles in the Understanding dimension ($D4_i$) determined based on the equation (7).

$$D4_i = \begin{cases} Se & if\ P4_i > 3 \\ G & if\ P4_i \leq 3 \end{cases} \tag{7}$$

with i is the dataset number i, Se : is a category of Sequential learning styles, G : is a Global learning style category, and P_4 : is the preference value at D_4 obtained from the equation (8).

$$P4_i = \frac{(ERH_i + LH_i))}{2}; \tag{8}$$

with the provision of:

$$ERH_i = \begin{cases} max(ER(i,:)) & if\ max(ER(i,:)) = ER(i,1) \\ & or\ max(ER(i,:)) = ER(i,2) \\ 0 & if\ max(ER(i,:)) = ER(i,3) \end{cases}$$

$$LH_i = \begin{cases} max(L(i,:)) & if\ max(L(i,:)) = L(i,2) \\ 0 & if\ max(L(i,:)) = L(i,1) \end{cases}$$

ERH and LH : is the maximum value of each preference in the Understanding dimension which successively is feature Exam_result (ER) and Length of stay in LMS features (L)

The sequence of dimensions obtained in each dataset number i using equations 1, 3, 5, and 7 that is $[D1_i, D2_i, D_i, D4_i]$ which then is used to identify learning style models that correspond to Table II. The initial centroids are taken based on the order of the FSLSM learning style model criteria in Table II which first was found in the dataset that the learning style model has been identified.

2) Calculating the distance of each dataset to the initial centroid and group the data into clusters with the closest centroid distance: This step is used to calculate the distance of data number i (x_i) to every initial centroid number k (c_k) using the Euclidean distance formula, as shown in the equation (9).

$$d_{ik} = \sqrt{\sum_{j=1}^{m} (x_{ij} - c_{kj})^2}. \tag{9}$$

where d_{ik} distance of data i to centroid on cluster k, $i = 1, 2, \ldots, n$ with n is the number of datasets, $k = 1, 2, \ldots, 16$, and m are the number of features.

Furthermore, group the data into clusters with the shortest distance. A data will be a member of the cluster k if the distance of the data to the centroid k is minimal, compared to those of other centroids. This can be calculated using equations (10).

$$c_i = Min(d_{ik}). \tag{10}$$

where c_i is the minimum cluster distance in each data point, then the new cluster membership is determined based on centroid with minimal distance.

3) Calculating a new centroid: This step is used to calculate the value of the new centroid by finding out the average value of data sets that become the members of the cluster using equation (11).

$$c_{kj} = \frac{\sum_{i=1}^{p} x_{ij}}{p}. \tag{11}$$

where p is the amount of members in the cluster k.

4) Calculating the distance of each data set to a new centroid and calculating the objective function values: This step is used to group data into clusters with the shortest distance using the new centroid generated in step 3, then calculated FO's value. The calculation value of FO is obtained from the closest distance from the new centroid between each data, which matches the cluster results from the previous iteration.

5) Determining the converging conditions of the iteration process: This step is employed to determine whether the iteration has converged or further iteration is required. The K-Means algorithm in this study was considered convergent if it fulfilled the following two conditions:

- The value of $Delta$ smaller than the threshold value (T) desired. The value of $Delta$ is the deviation of FO on two consecutive iterations, which can be calculated using equation (12).

  $$Delta = abs(FO_B - FO_L) \tag{12}$$

  with FO_B as the new value of FO and FO_L as the old value of FO. If the new iteration is done once, then FO_L can be given a reasonably sizeable initial value.

- There is no change in cluster membership.

C. Classification Process Using the Naive Bayesian Algorithm

Naive Bayesian (NB) is the algorithm assumes there is no correlation between variables for a given output value. The NB method is based on Bayes's Theorem. If there are two separate events X and K, then Bayes' Theorem is formulated using equation (13).

$$P\left(K|X\right) = \frac{P\left(X|K\right)}{P\left(X\right)}.P\left(K\right) \tag{13}$$

with:
X : Data with unknown class
K : Data hypothesis is a specific class
$P(K|X)$: Hypothesis probability K based on condition X
$P(K)$: Hypothesis probability K
$P(X|K)$: Probability X is based on a hypothesis K
$P(X)$: Probability X.

NB theorem is a classification process that requires some clues to determine the appropriate class for the sample being analyzed. Based on Bayes's Theorem in equation (13), the NB theorem can be formulated using equation (14).

$$P\left(K|F_1, \ldots, F_n\right) = \frac{P\left(F_1, \ldots, F_n|K\right)}{P\left(F_1, \ldots, F_n\right)}.P\left(K\right) \tag{14}$$

where, K represents class, while variable $F_1, \ldots, F_n$ represents the clue features needed to classify. Equation (14) explains that the probability of entering a sample of certain characteristics in a class K (posterior), which can also be formulated using equation (15).

$$Posterior = \frac{prior \times likelihood}{evidence} \tag{15}$$

with prior is the opportunity for class K to emerge before the entry of the sample, likelihood is the opportunity for the appearance of sample characteristics in the category K, and evidence is an opportunity for the emergence of sample characteristics globally.

Evidence values are always fixed for each class in one sample. The value of the posterior will later be compared with the values of the other class posterior to determine the sample that will be classified into the appropriate class. Further elaboration of the NB formula is done by explaining it $(K, F_1, \ldots, F_n)$ by using very high (naive) dependency assumptions. Each feature $(F_1, F_2, \ldots, F_n)$ is assumed to be independent of each other, so that equation (16) applies.

$$P\left(F_i|F_j\right) = \frac{P\left(F_i \cap F_j\right)}{P\left(F_j\right)} = \frac{P\left(F_i\right).P\left(F_J\right)}{P\left(F_J\right)} = P\left(F_i\right) \tag{16}$$

for $i \neq j$ can be formulated using equation (17).

$$P\left(F_i|K, F_j\right) = P\left(F_i|K\right) \tag{17}$$

or it can be written with notation as in equation (18).

$$P\left(k|F_1, F_2, F_3, \ldots, F_n\right) = P\left(K\right)\prod_{i=1}^{n} P\left(F_i|K\right) \tag{18}$$

Based on equation (18) the NB theorem for the classification process can be formulated using equation (19).

$$P\left(k|F\right) = P\left(F_i|k\right).P\left(F_2|k\right).P\left(F_3|k\right).\ldots.P\left(F_n|k\right). \tag{19}$$

D. Model Testing

A test to recognize the performance of the developed method consists of two processes, developed method they are clustering algorithm validation test and classification algorithm test. Cluster validity is obtained by measuring the cluster result based on a specific criteria. Cluster validity methods that are often used include Davies-Bouldin Index (DBI), Silhouette Index (SI), and Dunn Index (IN). Cluster validity measure used in this study is DBI since DBI has a reasonably good performance, which shows high accuracy and low time complexity [45].

David L. Davies and Donald W. Bouldin (1979) introduced the DBI matrix used to evaluate clusters. Cluster results are said to be good if the value of DBI is as small as possible (non-negative ≥ 0). Validity is done to measure how well the clustering is done by calculating the quantity and derivative features of a data set based on cohesion and separation values. The cohesion matrix or Sum of Square within-cluster (SSW) in the i cluster is formulated by the equation (20) [45].

$$SSW_i = \frac{1}{m_i}\sum_{j=1}^{m_i} d\left(x_j, c_j\right) \tag{20}$$

where m_i is the number of data in the cluster i, c_i is the centroid of the cluster i, and $d(x_i, c_j)$ is the same distance equation formula used when clustering process was performed Euclidean equation, city-block, and so on.

The matrix for separation between two clusters, for example, cluster number i and j using the formula Sum of Square Between Clusters (SSB) by measuring centroid distances c_i and c_j as shown in equation (21).

$$SSB_{i,j} = d\left(c_i, c_j\right) \tag{21}$$

Further, The value of DBI is obtained from equation (22).

$$DBI = \frac{1}{K}\sum_{i=1}^{K} max\left(R_{i,j}\right) \tag{22}$$

where K is the number of clusters and $R_{i,j}$ is the ratio of the total of sum of square within cluster for each corresponding cluster to their sum of square between clusters which is formulated using equation (23).

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \tag{23}$$

Testing the classification algorithm in this study in this conducted using a multi-class confusion matrix [46] $n \times n$ with $n = 16$, because it is used to analyze the classification of learning style detection containing 16 classes. If using a multi-class confusion matrix, the total number of false negatives (TFN), false positives (TFP), and true negative (TTN) for each class number i will be calculated based on Generalized (24), (25), and (26). equations. Total true positive (TTP (all)) in the system is obtained through equation (27).

$$TFN(i) = \sum_{j=1, j\neq i}^{n} f_{ij}. \tag{24}$$

$$TFP(i) = \sum_{j=1, j\neq i}^{n} f_{ji}. \tag{25}$$

$$TTN(i) = \sum_{j=1, j\neq i}^{n}\sum_{k=1, k\neq i}^{n} f_{jk} \tag{26}$$

$$TTP(all) = \sum_{j=1}^{n} f_{jj} \tag{27}$$

The performance of the proposed system in obtaining the relevant data is measured using Precision (P) or also called positive predictive value, while Recall (R) is used to measure the performance of the proposed classification in getting the relevant data to read. The class i used to calculate P and R for each class i equations (28) and (29).

$$P_i = \frac{TTP(all)}{TTP(all) + TFP(i)} \times 100\%. \tag{28}$$

$$R_i = \frac{TTP(all)}{TTP(all) + TFN(i)} \times 100\%. \tag{29}$$

The values of P and R are combined into one matrix called F-measure (F). The F is an average value of weighted harmonic between P and R. The F is calculated using equation (30).

$$F = 2 \times \frac{precision \times recall}{precision + recall} \times 100\%. \tag{30}$$

The performance of the proposed model built by the classification algorithm can be done by calculating the accuracy. The accuracy is calculated using the following equation (31).

$$Overall\_accuracy = \frac{TTP(all)}{The\_total\_amount\_of\_test\_data} \times 100\%. \tag{31}$$

IV. Analysis and Discussion

The results of the pre-processing process are obtained base on the data from PPG SPADA participants from Kemenristekdikti teachers teaching English subjects containing 500 data. The data set consists of 29 features, which consist of 9 features to determine the Processing dimension, 9 features for the Perception dimension, 6 features for the Input dimension, and 5 features for the Understanding dimension. Each feature contains several activities in each learning module consisting of 6 modules. The performance analysis of the proposed learning style detection model was tested using the Matlab R2013a application.

A. Results of Clustering Using Modified K-Means Algorithm

Based on the data set of PPG SPADA participants from Kemenristekdikti teachers teaching English subjects, initial centroid data is obtained from equations $(1) - (8)$ as shown in Table III.

TABLE III. Initial centroid data set

Centroid	Feature																												
	F1	F2	F3	E1	E2	E3	C1	C2	C3	R1	R2	R3	A1	A2	A3	Ex1	Ex2	Ex3	I1	I2	I3	M1	M2	M3	Er1	Er2	Er3	L1	L2
1	2	4	0	1	5	0	3	2	1	0	5	1	3	3	0	6	0	0	1	1	4	0	6	0	0	6	0	1	5
2	4	1	1	0	4	2	4	1	1	0	4	2	5	0	1	6	0	0	3	3	0	5	1	0	0	6	0	4	2
3	1	5	0	0	4	2	0	6	0	6	0	0	1	5	0	5	0	1	4	2	0	0	2	4	0	4	2	1	5
4	4	1	1	5	0	1	0	5	1	3	1	2	5	1	0	2	3	1	4	2	0	3	3	0	2	1	3	1	5
5	0	6	0	2	3	1	3	1	2	1	0	5	1	4	1	3	0	3	0	5	1	3	2	1	2	3	1	2	4
6	3	3	0	3	1	2	0	5	1	3	1	2	3	2	1	0	0	6	0	6	0	3	1	2	0	1	5	1	5
7	0	5	1	4	1	1	1	5	0	2	4	0	0	0	6	2	2	2	4	0	2	5	0	1	3	0	3	1	5
8	2	1	3	0	5	1	0	5	1	3	2	1	2	2	2	1	0	5	6	0	0	4	0	2	5	0	1	4	2
9	0	2	4	1	0	5	3	1	2	1	3	2	0	6	0	3	0	3	0	0	6	1	3	2	1	5	0	0	6
10	1	3	2	1	0	5	6	0	0	1	4	1	4	1	1	6	0	0	1	5	0	2	4	0	0	0	6	5	1
11	3	1	2	3	0	3	0	1	5	4	0	2	2	4	0	3	3	0	5	0	1	1	1	4	2	3	1	0	6
12	2	3	1	3	1	2	0	3	3	2	4	0	4	0	2	0	6	0	2	0	4	2	1	3	2	1	3	0	6
13	0	1	5	0	5	1	1	3	2	1	4	1	1	4	1	0	1	5	0	4	2	1	5	0	4	2	0	3	3
14	2	1	3	3	2	1	2	1	3	2	1	3	4	2	0	4	1	1	1	5	0	5	0	1	0	2	4	0	6
15	0	2	4	2	1	3	1	0	5	1	5	0	2	0	4	1	3	2	3	3	0	3	0	3	0	3	3	2	4
16	1	2	3	3	1	2	3	3	0	1	3	2	2	2	2	3	1	2	3	1	2	2	4	0	2	0	4	4	2

Test results using $Threshold(T) = 0.1$ produced the clustering results that reached a convergent condition in the 18 iteration with the clustering results, as shown in Table IV.

TABLE IV. Results of clustering dataset

Cluster	Learning Style Model	Count
1	(A,S,Vi,Seq)	4
2	(A,S,Vi,G)	17
3	(A,S,Ve,Seq)	4
4	(A,S,Ve,G)	18
5	(A,I,Vi,Seq)	65
6	(A,I,Vi,G)	10
7	(A,I,Ve,Seq)	17
8	(A,I,Ve,G)	18
9	(R,S,Vi,Seq)	9
10	(R,S,Vi,G)	4
11	(R,S,Ve, Seq)	28
12	(R,S,Ve,G)	75
13	(R,I,Vi,Seq)	39
14	(R,I,Vi,G)	18
15	(R,I,Ve,Seq)	79
16	(R,I,Ve,G)	95
	Total	500

Based on Table IV, the cluster with the most members is the cluster with Reflective, Intuitive, Verbal, and Global learning styles. This shows that the participant of the e-learning platform involved in this study are mostly in reflective observation learning style type that prefers to think for themselves solving problems that are calmly faced first. Participants also prefer innovation and do not like lectures that involve memorization and routine calculations. Participants also prefer to get information from discussions and learn effectively by explaining to others. Furthermore, participants in this group prefer to receive random material, so that they can solve complex problems quickly when they get the big picture.

Testing the validity of the clustering algorithm is carried out by comparing the maximum value of $DBI(R)$ in each cluster between the modified algorithm with the original K-Means. The value of R in each group for one experiment is depicted in Fig. 3.

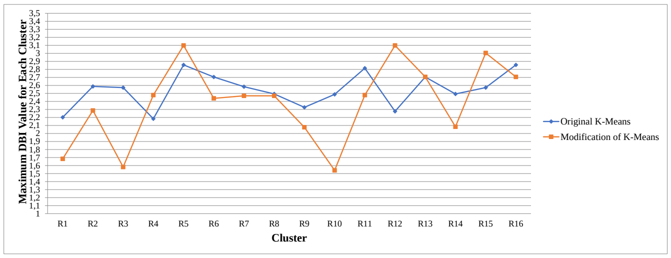


Fig. 3. Comparison of the value of R on the original K-Means algorithm with the modified K-Means

Based on Fig. 3, the average value of R for the modified K-Means algorithm is smaller than the original K-Means. Apart from that, DBI value of the modified K-Means algorithm is 2.39 lower than the original K-Means i.e., 2.55. Based on 15 repetition, the modified K-Means algorithm also shows stable DBI and R values compared to the original K-Means, which fluctuates in each experiment, as shown in Fig. 4.

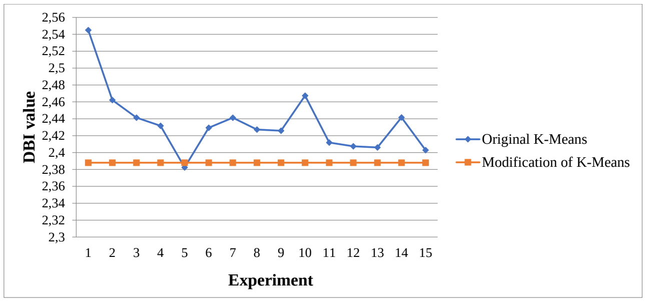


Fig. 4. The results of testing the value of DBI on the original K-Means algorithm with K-Means modification

Fig. 4 shows that value of DBI for the original K-Means Algorithm is unstable, and the clustering result for each data set also differs for each attempt. This is because the value of initial centroid always changes since it is determined randomly, which causes the validity of the algorithm always to improve. While the value of DBI for K-Means algorithm that had been modified remains similar to the clustering result for each data set, also it does not show any change. This result shows that the modified K-Means algorithm is good enough compared to the original K-means so that the data of the clustering result using K-Means algorithm that have been modified increases the performance of the classification algorithm to detect the learning style of the participants of PPG SPADA Ristekdikti of the English teachers.

B. Classification Results using the Naive Bayesian Algorithm

Based on the test results from 500 data sets between class labels, the results of clustering using the modified K-Means algorithm that 358 out of 500 data $(71, 60\%)$ have predicted classes that equal to the correct class. In contrast, the class labels that are different from the prediction results are 142 data $(28, 40\%)$. The precision and recall values are shown in Table V.

TABLE V. The value of precision and recall in each class

Class	**Description Learning Style Model**	**Precision (P) (%)**	***Recall (R)* (%)**
1	(A,S,Vi,Seq)	40.00	100.00
2	(A,S,Vi,G)	44.44	70.59
3	(A,S,Ve,Seq)	100.00	100.00
4	(A,S,Ve,G)	45.45	83.33
5	(A,I,Vi,Seq)	83.05	75.38
6	(A,I,Vi,G)	69.23	90.00
7	(A,I,Ve,Seq)	56.52	76.47
8	(A,I,Ve,G)	51.72	83.33
9	(R,S,Vi,Seq)	81.82	100.00
10	(R,S,Vi,G)	100.00	100.00
11	(R,S,Ve, Seq)	63.64	50.00
12	(R,S,Ve,G)	85.00	45.33
13	(R,I,Vi,Seq)	84.38	69.23
14	(R,I,Vi,G)	80.00	88.89
15	(R,I,Ve,Seq)	71.23	65.82
16	(R,I,Ve,G)	81.00	85.26
	Average	71.09	80.23

Table V shows the average P is 71.09%, which means that the level of accuracy of the detection information of the learning style model desired by the user with the answers given by the proposed model is quite high. While the average value of R is 80.23%, which shows the performance of the proposed model is quite good, above 70%. Table V also shows 12 of the 16 class learning style models have value R higher than 70%, which means 75% of learning style of the course participants were successfully detected using a combination of modified K-Means algorithm with NB classification.

The proposed method successfully classifies each FSLSM learning style model quite well. This can be seen from the average value of precision and recall, which is almost balanced, and the F-Measure value is 75.38%, which is higher than 70%. The accuracy of the proposed model is also quite good, which is 71.6%. This shows the level of similarity of the prediction of the learning styles of PPG SPADA participants of the Ministry of Research, Technology, and Higher Education teachers of English subjects, and the learning styles model is quite close.

The performance of the proposed automatic learning style detection model is compared to the learning style detection model if the clustering algorithm uses the original K-Means algorithm performed by measuring the average value P, R, accuracy value, and F-Measures tested 10 times. The test results are shown in Tables VI, VII, VIII, and IX. Based on the results of testing the average value of precision, recall, accuracy, and F-Measure as shown in Tables VI, VII, VIII, and IX can be seen that the use of a modified of the K-Means algorithm to form labels before classification has increased. This shows that changes made to the K-Means algorithm improves the performance of the learning style detection model when using the original K-Means algorithm. In addition to the increasing performance of the proposed method, the average values of precision, recall, accuracy, and F-Measure also did not change. It shows if the performance of the technique of learning style detection proposed has stable performance.

TABLE VI. Comparison of the results of testing the average precision of the proposed learning style detection model using a modification of the K-Means algorithm with the original K-Means

Clustering	**The value of Precision in each Experiment (%)**									
Algorithm	1	2	3	4	5	6	7	8	9	10
Original K-Means	64.21	67.27	61.93	68.26	65.31	67.22	65.48	68.51	61.43	64.55
Modification of K-Means	71.09	71.09	71.09	71.09	71.09	71.09	71.09	71.09	71.09	71.09

TABLE VII. Comparison of the average recall results of the proposed learning style detection model using a modification of the K-Means algorithm with the original K-Means

Clustering	**The value of Recall in each Experiment (%)**									
Algorithm	1	2	3	4	5	6	7	8	9	10
Original K-Means	69.75	76.76	72.71	73.57	74.18	72.01	70.03	73.7	66.8	71.41
Modification of K-Means	80.23	80.23	80.23	80.23	80.23	80.23	80.23	80.23	80.23	80.23

TABLE VIII. Comparison of the average test results for the accuracy of the proposed learning style detection model using a modification of the K-Means algorithm with the original K-Means

Clustering	**The value of accuracy in each Experiment (%)**									
Algorithm	1	2	3	4	5	6	7	8	9	10
Original K-Means	64.80	69.60	62.60	69.80	67.40	66.00	65.20	69.80	69.20	66.20
Modification of K-Means	71.60	71.60	71.60	71.60	71.60	71.60	71.60	71.60	71.60	71.60

TABLE IX. Comparison of average F-Measure test results for proposed learning style detection models using modification of the K-Means algorithm with the original K-Means

Clustering	**The value of F-Measure in each Experiment (%)**									
Algorithm	1	2	3	4	5	6	7	8	9	10
Original K-Means	66.87	71.71	66.89	70.82	69.47	69.54	67.68	71.01	63.67	67.81
Modification of K-Means	75.38	75.38	75.38	75.38	75.38	75.38	75.38	75.38	75.38	75.38

V. Conclusion

This research succeeded in building an automatic learning style detection model using a combination of K-Means algorithm modification with Naive Bayesian. Based on the test results, there is a modification of the K-Means algorithm, which is used to form labels on the learning force detection models proposed in this study can improve the performance of grouping the data sets when compared to the original K-Means algorithm. The results of testing the validity of the modified K-Means algorithm are better than the original K-Means algorithm. Besides that, the DBI value on the modified K-Means algorithm has the same value every time it is implemented. This shows that the modification of the K-Means algorithm is more stable than the original K-Means algorithm so that the labels of each data set do not change.

The proposed learning style detection model by using a combination of modification of the K-Means algorithm before classification can improve the performance of the learning style detection model if the labeling process uses the original K-Means algorithm. The average precision and recall values of the test data set are 71.09% and 80.23%, which means the proposed model for detecting learning styles works well. The accuracy value of the proposed model is still quite good, i.e., 71.6%, which is higher than the average accuracy of the learning style detection model that uses the original K-Means algorithm for the clustering process, which is 64.8%. This shows that the level of closeness between predictions with the original learning style model is quite high.

As part of future work, the proposed model allows for increased accuracy, precision, and recall values by improving the performance of the Naive Bayesian classification method using the Augmented Naive Bayesian Tree algorithm or Artificial Neural Network-based classification algorithm.

References

[1] M. P. P. Liyanage, K. S. L. Gunawardena, and M. Hirakawa, "Using Learning Styles to Enhance Learning Management Systems," *International Journal on Advances in ICT for Emerging Regions 2014*, vol. 07, no. 02, pp. 1–10, 2014.

[2] S. Graf, T.-C. Liu, and Kinshuk, "Analysis of learners' navigational behaviour and their learning styles in an online course," *Journal of Computer Assisted Learning*, vol. 26, no. 2, pp. 116–131, Mar 2010.

[3] J. Feldman, A. Monteserin, and A. Amandi, "Automatic detection of learning styles: state of the art," *Artificial Intelligence Review*, vol. 44, no. 2, pp. 157–186, Aug 2015.

[4] O. ZINE, A. DEROUICH, and A. TALBI, "A Comparative Study of the Most Influential Learning Styles used in Adaptive Educational Environments," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 11, pp. 520–528, 2019.

[5] M. P. P. Liyanage, K. L. Gunawardena, and M. Hirakawa, "Detecting Learning Styles in Learning Management Systems Using Data Mining," *Journal of Information Processing*, vol. 24, no. 4, pp. 740–749, 2016.

[6] T. Sheeba and R. Krishnan, "Prediction of student learning style using modified decision tree algorithm in e-learning system," in *Proceedings of the 2018 International Conference on Data Science and Information Technology - DSIT '18*. New York, USA: ACM Press, 2018, pp. 85–90.

[7] M. P. Pitigala Liyanage, K. S. Gunawardena, and M. Hirakawa, "A framework for adaptive learning management systems using learning styles," in *2013 International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, Dec 2013, pp. 261–265. [Online]. Available: https://ieeexplore.ieee.org/document/6761188/

[8] L. X. Li and S. S. Abdul Rahman, "Students' learning style detection using tree augmented naive Bayes," *Royal Society Open Science*, vol. 5, no. 7, pp. 1–13, Jul 2018.

[9] S. R. Viola, S. Graf, Kinshuk, and T. Leo, "Investigating relationships within the Index of Learning Styles: a data driven approach," *Interactive Technology and Smart Education*, vol. 4, no. 1, pp. 7–18, Feb 2007.

[10] C. C. Hosford and W. A. Siders, "Felder-Soloman's Index of Learning Styles: Internal Consistency, Temporal Stability, and Factor Structure," *Teaching and Learning in Medicine*, vol. 22, no. 4, pp. 298–303, Oct 2010.

[11] W. Jingyun and M. Takahiko, "The Reliability and Validity of Felder-Silverman Index of Learning Styles in Mandarin Version," *Information Engineering Express International Institute of Applied Informatics*, vol. 1, no. 3, pp. 1–8, 2015.

[12] A. S. Ovariyanti and H. B. Santoso, "An adaptation of the Index of Learning Style (ILS) to Indonesian version: A contribution to Index of Learning Style (ILS), validity and reliability score," in *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE, Oct 2016, pp. 129–134.

[13] E. Švarcová and K. Jelínková, "Detection of Learning Styles in the Focus Group," *Procedia - Social and Behavioral Sciences*, vol. 217, pp. 177–182, 2016.

[14] P. García, A. Amandi, S. Schiaffino, and M. Campo, "Using Bayesian Networks to Detect Students' Learning Styles in a Web-based education system," in *Proc of ASAI, Rosario*, 2005, pp. 115–126.

[15] N. Ahmad, Z. Tasir, J. Kasim, and H. Sahat, "Automatic Detection of Learning Styles in Learning Management Systems by Using Literature-based Method," *Procedia - Social and Behavioral Sciences*, vol. 103, pp. 181–189, Nov 2013.

[16] A. S. M. Ghazali, S. F. M. Noor, and S. Saad, "Review of personalized learning approaches and methods in e-learning environment," *Proceedings - 5th International Conference on Electrical Engineering and Informatics: Bridging the Knowledge between Academic, Industry, and Community, ICEEI 2015*, pp. 624–627, 2015.

[17] E. S. Amir, M. Sumadyo, D. I. Sensuse, Y. G. Sucahyo, and H. B. Santoso, "Automatic detection of learning styles in learning management system by using literature-based method and support vector machine," *2016 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2016*, pp. 141–144, 2017.

[18] J. Bernard, T.-W. Chang, E. Popescu, and S. Graf, "Learning style Identifier: Improving the precision of learning style identification through computational intelligence algorithms," *Expert Systems with Applications*, vol. 75, pp. 94–108, Jun 2017.

[19] O. El Aissaoui, Y. El Madani El Alami, L. Oughdir, and Y. El Allioui, "Integrating web usage mining for an automatic learner profile detection: A learning styles-based approach," in *2018 International Conference on Intelligent Systems and Computer Vision (ISCV)*, vol. 2018-May. IEEE, Apr 2018, pp. 1–6.

[20] I. Karagiannis and M. Satratzemi, "An adaptive mechanism for Moodle based on automatic detection of learning styles," *Education and Information Technologies*, vol. 23, no. 3, pp. 1331–1357, May 2018.

[21] R. R. M. III, M. A. Ballera, S. C. Ambat, and M. F. Dumlao, "Comparative Analysis of Data Mining Techniques for Classification of Student ' s Learning Styles," *International Conference on Advances in Science, Engineering and Technology (ICASET-17) Sept.*, pp. 65–70, Sep 2017.

[22] P. García, A. Amandi, S. Schiaffino, and M. Campo, "Evaluating Bayesian networks' precision for detecting students' learning styles," *Computers & Education*, vol. 49, no. 3, pp. 794–808, Nov 2007.

[23] M. Abdullah, A. Alqahtani, J. Aljabri, R. Altowirgi, and R. Fallatah, "Learning Style Classification Based on Student ' s Behavior in Moodle Learning Management System," *Transactions on Machine Learning and Artificial Intel lntelligence*, vol. 3, no. 1, pp. 28–40, Feb 2015.

[24] L. Mahnane and M. Hafidi, "Automatic detection of learning styles based on dynamic Bayesian network in adaptive e-learning system," *International Journal of Innovation and Learning*, vol. 20, no. 3, pp. 289–308, 2016.

[25] O. E. Aissaoui, Y. E. A. EL Madani, L. OUGHDIR, and Y. E. Allioui, "Combining supervised and unsupervised machine learning algorithms to predict the learners' learning styles," *Procedia Computer Science*, vol. 148, no. February, pp. 87–96, 2019.

[26] A. Kika, L. Leka, S. Maxhelaku, and A. Ktona, "Using Data Mining Techniques on Moodle Data For Classification Of Student's Learning Styles," in *Proceedings of the 47th International Academic Conference, Prague*. International Institute of Social and Economic Sciences, 2019, pp. 26–33.

[27] N. Ahmad, Z. Tasir, and N. A. Shukor, "Using Automatic Detection to Identify Students' Learning Style in Online Learning Environment – Meta Analysis," in *2014 IEEE 14th International Conference on Advanced Learning Technologies*. IEEE, Jul 2014, pp. 126–130.

[28] A. EL Mezouary, B. Hmedna, and O. Baz, "An evaluation of learner clustering based on learning styles in MOOC course," in *2019 International Conference of Computer Science and Renewable Energies (ICCSRE)*. IEEE, Jul 2019, pp. 1–5.

[29] D. M. Farid, L. Zhang, C. M. Rahman, M. Hossain, and R. Strachan, "Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1937–1946, Mar 2014.

[30] P. Fränti and S. Sieranoja, "How much can k-means be improved by using better initialization and repeats?" *Pattern Recognition*, vol. 93, pp. 95–112, Sep 2019.

[31] S. Penger, "Education In Slovenia : An Experiential," *International Business*, vol. 7, no. 12, pp. 25–44, 2008.

[32] P. Sangvigit, "Correlation of Honey & Mumford Learning Styles and Online Learning media preference," *International Journal of Computer Technology and Applications*, vol. 3, pp. 1312–1317, 2012.

[33] M. Dziedzic, F. B. de Oliveira, P. R. Janissek, and R. M. Dziedzic, "Comparing learning styles questionnaires," in *2013 IEEE Frontiers in Education Conference (FIE)*. IEEE, Oct 2013, pp. 973–978.

[34] A. Gogus and G. Ertek, "Learning and Personal Attributes of University Students in Predicting and Classifying the Learning Styles: Kolb's Nine-region Versus Four-region Learning Styles," *Procedia - Social and Behavioral Sciences*, vol. 217, pp. 779–789, Feb 2016.

[35] S. Rajper, N. A. Shaikh, Z. A. Shaikh, and G. Ali Mallah, "Automatic Detection of Learning Styles on Learning Management Systems using Data Mining Technique," *Indian Journal of Science and Technology*, vol. 9, no. 15, pp. 1–5, May 2016.

[36] A. A. Kalhoro, S. Rajper, and G. A. Mallah, "Detection of E-Learners' Learning Styles: An Automatic Approach using Decision Tree," *International Journal of Computer Science and Information Security*, vol. 14, no. 8, pp. 420–425, 2016.

[37] M. P. Hamzah, W. Fatin, F. Yahya, N. Maizura, and M. Noor, "Learning Style Detection By Using Literature-Based Approach : A Conceptual Design," *International Symposium on Research in Innovation and Sustainability 2014 (ISoRIS '14) 15-16 October 2014, Malacca, Malaysia*, vol. 2014, pp. 1493–1497, 2014.

[38] O. Pantho, "Using Decision Tree C4 . 5 Algorithm to Predict VARK Learning Styles," *International Journal of the Computer, the Internet and Management*, vol. 24, no. 2, pp. 58–63, 2016.

[39] M. S. Hasibuan, L. E. Nugroho, and P. I. Santosa, "Model detecting learning styles with artificial neural network," *Journal of Technology and Science Education*, vol. 9, no. 1, pp. 85–95, Feb 2019.

[40] S. Sweta and K. Lal, "Learner Model for Automatic Detection of Learning Style Using FCM in Adaptive E-Learning System," *IOSR Journal of Computer Engineering*, vol. 18, no. 2, pp. 18–24, 2016.

[41] H. D. Herman Dwi, "The Evaluation of a Moodle Based Adaptive e-Learning System," *International Journal of Information and Education Technology*, vol. 4, no. 1, pp. 89–92, 2014.

[42] I. Azzi, A. Jeghal, A. Radouane, A. Yahyaouy, and H. Tairi, "A robust classification to predict learning styles in adaptive E-learning systems," *Education and Information Technologies*, vol. 1, pp. 772–786, Aug 2019.

[43] D. El-Hmoudova, "Self-efficacy for Learning vs ILS Results in a group of English Learning Bachelor Students," *Procedia - Social and Behavioral Sciences*, vol. 199, pp. 563–570, Aug 2015.

[44] L. D. Ferreira, G. Spadon, A. C. Carvalho, and J. F. Rodrigues, "A comparative analysis of the automatic modeling of Learning Styles through Machine Learning techniques," *Proceedings - Frontiers in Education Conference, FIE*, vol. Oct, pp. 1–8, 2019.

[45] S. Petrovic, "A Comparison Between the Silhouette Index and the Davies-Bouldin Index in Labelling IDS Clusters," in *11th Nordic Workshop on Secure IT-systems*, 2006, pp. 53–64.

[46] C. Manliguez, "Generalized Confusion Matrix for Multiple Classes," Tech. Rep., 2016.

Improved Candidate Generation for Pedestrian Detection using Background Modeling in Connected Vehicles

Ghaith Al-refai[1], Osamah A. Rawashdeh[2]
School of Electrical and Computer Engineering
Oakland University, Rochester, Michigan 48309

***Abstract*—Pedestrian detection is widely used in today's vehicle safety applications to avoid vehicle-pedestrian accidents. The current technology of pedestrian detection utilizes onboard sensors such as cameras, radars, and Lidars to detect pedestrians, then information is used in a safety feature like Automatic Emergency Braking (AEB). This paper proposes pedestrian detection system using vehicle connectivity, image processing and computer vision algorithms. In the proposed model, vehicles collect image frames using on-vehicle cameras, then frames are transferred to the Infrastructure database using Vehicle to Infrastructure communication (V2I). Image processing and machine learning algorithms are used to process the infrastructure images for pedestrian detection. Background modeling is used to extract the foreground regions in an image to identify regions of interest for candidate generation. This paper explains the algorithms of the infrastructure pedestrian detection system, which includes image registration, background modeling, image filtering, candidate generation, feature extraction, and classification. The paper explains the MATLAB implementation of the algorithm with a road-collected dataset and provides analysis for the detection results with respect to detection accuracy and runtime. The algorithm implementation results show an improvement in the detection performance and algorithm runtime.**

Keywords*—Pedestrian detection; computer vision; image processing; machine learning; vehicle safety

I. Introduction

Between 2010 and 2013 the number of registered vehicles increased by 16% [1]. This causes a significant increase in the number of road accidents and road fatalities. The number of worldwide deaths because of road accidents was 1.25 million in 2015 [1]. Many safety solutions have been introduced in vehicles to improve road safety: Advanced Driver Assistance Systems (ADAS) is one of them. ADAS technology utilizes on-vehicle sensors to detect surrounding objects and then analyze detection results to avoid accidents and drive safely. Radar, Lidar, and ultrasonic sensors are examples of sensors that are used in ADAS. Cameras are a widely used sensor in ADAS due to the low cost and the rich information they provide. Image processing and machine learning are used to detect objects of interest in image frames, and the results are used in many safety features. A basic vision-based object detection system includes the following processes: image acquisition using a camera, candidate generation for the object of interest, feature extraction to describe the candidates, and finally, a trained classifier to classify candidates.

The candidate generation process is a very critical step in the detection system and it has a direct impact on the detection accuracy and the processing requirements. There are many approaches for pedestrian candidate generation. The basic approach is the multiple size image scanning, where the whole image is scanned by a sliding window at multiple sizes to detect pedestrians at different sizes and distances. Papageorgiou and Broggi used a window of 64x128 for pedestrian detection and image sizing between 0.2 to 2 of its original size with a step of 0.1 [2]. The flat world approach for candidate generation assumes the world is flat, and it generates the candidates from the ground plane level [3]. This approach provides inaccurate results when the camera location changes with respect to the ground because of vehicle dynamics and road slope. Many solutions introduced to stabilize the images using horizontal edges histogram [4] and features matching [5], but they are computationally expensive. The stereo vision is another approach for candidate generation, where a constructed 3-D map is used to identify the regions of interest to generate the candidates [6]. This approach is expensive since it requires two cameras for the 3-D map construction and it requires a lot of computations.

The current on-board candidate generation approaches can't distinguish between static and moving objects in an image. This leads to the generation of many unnecessary candidates, which can cause false detection and increases the algorithm runtime. An example of this is generating candidates for trees and buildings in an image and misclassifying them as pedestrians.

This paper introduces a new model for candidate generation using connected vehicles and background modeling. The model suggests that images of roads are collected by on-vehicle cameras and the frames are transferred to the infrastructure using V2I. Images that belong to the same location are processed together to generate a background model and improve candidate generation and then pedestrian detection system.

According to a study done by National Highway Safety Admiration (NHTSA), V2V can address 79% of all vehicle crashes while V2I can handle 28% of traffic light accidents [7]. Because of connected vehicles potential in road safety, many researches have been aiming to extend connected vehicles, capabilities in images and video sharing. Video sharing using V2V was experimentally implemented in [8]. Vehicle connectivity for video sharing using 5G network was proved in [9].

This paper focuses on the image processing and machine learning algorithms that needs to be implemented in the infrastructure for accurate detection results. The second section of this paper provides an overview for the infrastructure pedestrian detection system. The third section explains the algorithms of the infrastructure detection system. The fourth section explains the infrastructure implementation in MATLAB. The fifth section shows the algorithm results and compares them to a reference on-board detection algorithm. The sixth section summarizes the conclusions of this research.

II. Infrastructure System Overview

Implementing a pedestrian detection system in the connected vehicles needs special requirements in the vehicle, V2I communication channel, and the infrastructure system. This section provides an overview of the infrastructure background modeling for pedestrian detection.

A. Vehicle Components

The system requires a vehicle with a forward-looking camera for video collection. V2I transceiver is also required to transfer image frames from vehicle to infrastructure. Other information such as GPS data and vehicle dynamics shall be transferred along with the images for registration.

B. V2I Communication

The image frames and their associated data are transferred via V2I channel. The channel shall have enough bandwidth for image transfer. The channel shall have acceptable latency for real time detection. The channel shall meet other communication specifications such as data encryption and data security.

C. Infrastructure Database

The image frames and their associated data are stored in the infrastructure database. The database is real time maintained with every passing vehicle. Image frames that belong to the same location are grouped together.

D. Infrastructure Pedestrian Detection System

The infrastructure has the history images of a location that was collected by the passed vehicles. History image availability makes pedestrian detection in the infrastructure different from onboard approaches. The infrastructure pedestrian detection system includes following processes:

1) Image registration: Vehicle cameras have different specifications such as resolution, field of view and orientation. Therefore, image registration is required to match images together to be processed as a group. There are many registration techniques to handle this challenge. Vitoza and Flusser provided a review for image registration approaches that can be utilized in this step [10]. Harris -Stephen approach is used in the pedestrian detection system for images alignment and registration.

2) Background modeling: Image frames belonging to the same location are used for background modeling. The background model is used for foreground pixels extraction from the current frame. There are many approaches for background modeling. The used background modeling shall have the ability to handle dynamic changes in background images, such as removing and inserting objects. The background model shall be real time maintained to have the latest updates of road conditions.

3) Foreground regions extraction and candidate generation: The background model is compared to the current image frame to extract the foreground pixels. Image filtering is required to remove the noise and construct the shape of the moving regions. Finally, candidates are generated only from the foreground regions by applying image thresholding.

4) Feature extraction and classification: Features such as edges, corners, and colors are extracted from the candidates for better object description. The feature vectors of the candidates are passed to a trained classifier to classify them as pedestrians or non-pedestrians. Gerónimo and López provided a review for the different approaches of feature extraction and classification in pedestrian detection [11]. Fig. 1 provides the block diagram of the infrastructure improved candidate generation in pedestrian detection using background modeling. Al-refai, Horani and Rawashdeh provided a detailed system architecture and specifications of the infrastructure pedestrian detection system [12].

III. Infrastructure Pedestrian Detection System Algorithms

This section introduces and explains the algorithms to implement the infrastructure pedestrian detection system. The proposed system includes image registration, background modeling, foreground regions extraction, image filtering, candidate generation, feature extraction, and classification.

Harris-Stephens approach for corner detection is used for image registration and matching. The Gaussian Mixture Model (GMM) is used for background modeling and maintenance. The foreground regions in images are extracted using the GMM model. The foreground digital mask is filtered using morphological filters. Candidates are generated from the moving regions in the foreground digital mask. Finally, Histogram Oriented Gradient (HOG) and Support Vector Machine (SVM) are used for candidate feature extraction and classification.

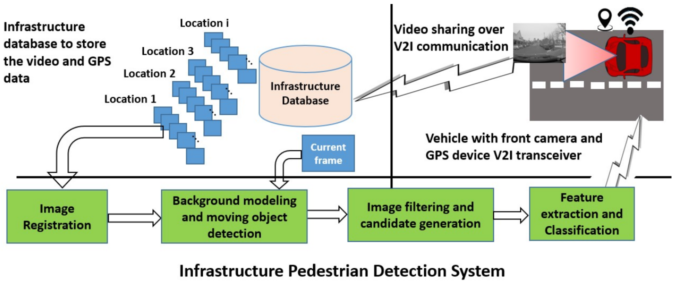

Infrastructure database to store the video and GPS data
Location i
Location 3
Location 2
Location 1
Infrastructure Database
Video sharing over V2I communication
Current frame
Vehicle with front camera and GPS device V2I transceiver
Image Registration
Background modeling and moving object detection
Image filtering and candidate generation
Feature extraction and Classification
Infrastructure Pedestrian Detection System

Fig. 1. Background modeling for improved candidate generation in pedestrian detection using V2I block diagram

The following subsections explain the algorithms and the mathematical model for each process.

A. Harris-Stephens Corner Detection for Image Registration

Images collected by vehicle's front cameras don't have the same specifications. They have different rotations, field of view and resolution. In this step, the images in the database for a certain location will be registered and aligned to a reference image. The reference image is selected to be the first image captured by a vehicle for the location. Reference image shall be updated every certain amount of time to include the latest updates in the scene. In our collected dataset, the first image in the sequence is selected to be the reference image.
Image registration includes three steps: Feature extraction from the reference image and the target image, Feature mapping and image transformation.

Harris-Stephen proposed an algorithm for corner detection [13]. This algorithm is used in our system for image registration. Corner features are selected as a control point in the registration for the following reasons:

- Corners are common features in roads
- Corners invariant to geometric changes
- Corners are invariant to resolution change
- Corners are Partially invariant to intensity values

Corners are detected by measuring the change in the intensity values of the pixels in the x and y directions. If the change in the intensity values are large in both directions, then it is considered as a corner. More information about the algorithm implementation can be found in [13].
The next step is to match the features in the reference image and the target image.One of the best algorithms for feature matching is the nearest neighbor distance ratio (NNDR) [14]. The NNDR algorithm works as following:

- Compute the distance between the corners vector in the reference image f_r and the nearest neighbor corners vector in the target image f_{t1} using the sum of square root differences (SSD).

$$d_1 = \sum_{i=1}^{n} (f_{t1} - f_r)^2 \tag{1}$$

where
L: The length of the feature vector i
f_r: A feature vector in the reference image
f_{t1}: The nearest neighbor vector in the target image

- Compute the distance between the reference image feature vector and the second nearest neighbor in the target image

$$d_2 = \sum_{i=1}^{n} (f_{t2} - f_r)^2 \tag{2}$$

- If the ratio between the two distances d_1/d_2 is low, then it is a good match. If the ration is greater than the threshold "MaxThrshld", then the algorithm eliminates the matched as ambiguous.

The last step of the registration is the image transformation. In this step the transformation factors are predicted. this includes image rotation in the pitch, yaw and roll directions, image translation and scaling. The transformation matrix is 3 x 3 with eight unknowns, so the minimum required matching points between the reference image and the target image shall be four. Fig. 2 shows the block diagram of the image registration block.

Fig. 3 shows an example of corner detection and feature matching for a rotated image. Registered images are passed to GMM for background modeling as explained in the next section.

B. GMM for Background Modeling and Foreground Extraction

This part explains GMM for background modeling and feature extraction and compares GMM to the mean filter for background modeling to highlight the advantages of GMM over the basic approaches for background modeling and foreground extraction.

Background modeling and foreground pixel extractions are generally done in three steps: background modeling, background maintenance, and foreground detection. The background modeling step uses the previous image frames to create a model of the background. The background model can

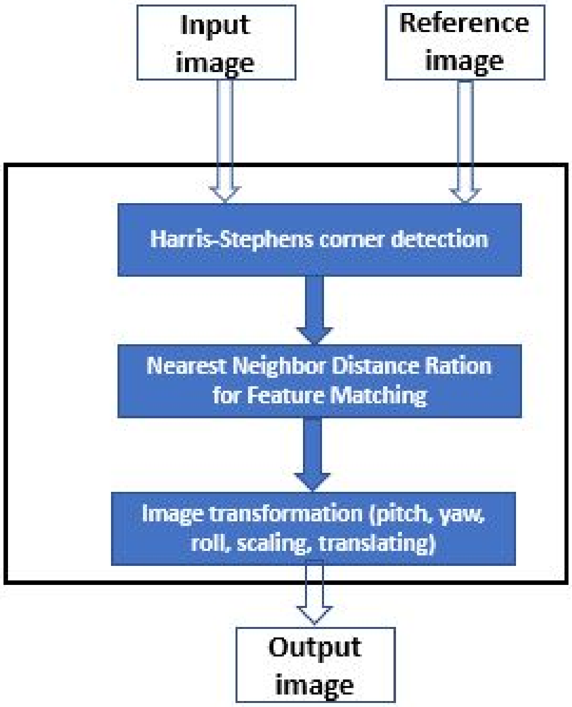

Input image
Reference image
Harris-Stephens corner detection
Nearest Neighbor Distance Ration for Feature Matching
Image transformation (pitch, yaw, roll, scaling, translating)
Output image

Fig. 2. The registration system block diagram

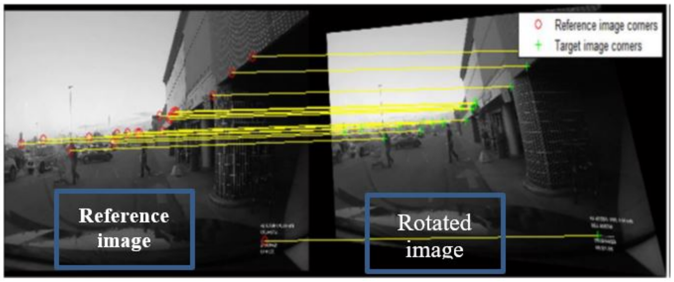


Fig. 3. Image registration using Harris-Stephens corner detection and nearest neighbor ratio for feature matching

be an image or mathematical function such as a probability density function.

Many changes occur in images for a location over time. As objects move, objects are removed from the background and others are inserted. Background maintenance is needed as a mechanism to adapt the background to the latest changes. Many approaches were developed for background maintenance, and they are generally categorized as a blind maintenance and a selective maintenance.

The maintained background model is used to extract the foreground pixels by comparing the current image to the background. The simplest approach to extract the foreground regions is to subtract the current frame from the background model. Other approaches use statistical modeling for background estimation.

One of the basic ways for background modeling is the mean filter [15], which is given by:

$$B(x,y,t) = \frac{1}{n}\sum_{i=1}^{n} I(x,y,t-1), \tag{3}$$

where B(x,y,t) is the background model at time t, I(x,y,t) is the image frame with (x,y) pixels at time t, and n is the total number of image frames. Then foreground pixels are determined by:

$$F(x,y,t) = |I(x,y,t) - B(x,y,t)| > T \tag{4}$$

where T is a fixed threshold value. Median filter is also used for background modeling [16]. The background is maintained by adding a portion of the current image to the background model:

$$B(x,y,t+1) = (1-\alpha)B(x,y,t) + \alpha I(x,y,t), \tag{5}$$

where α is the learning rate which is a constant in [0,1], usually it is 0.05.

Basic approaches have many problems in handling the dynamic changes in the background, such as light variations and shadowing. Also, the basic models require a large memory. Statistical approaches were introduced to handle the dynamic changes in the background. In the statistical approaches, the intensity values of the pixels are modeled in a probability density functions (PDF). Then the PDFs are used to estimate the current pixel as belonging to the background or not. Background modeling using a single Gaussian function is proposed in [17]. However, one PDF for each pixel is insufficient to model the background in a dynamic environment. To solve the problem, a mixture of Gaussians is used to model the background [18]. It is also called Gaussian Mixture Model (GMM). GMM solves many issues for background modeling such as removed background objects and inserted background objects. The memory requirement of GMM is less than the basic approaches. More details about background modeling approaches and foreground detection can be found [19] and [20].

GMM is used in our proposed model as introduced by Stauffer and Grimson [18]. A simplified explanation of GMM mathematical model is provided below:

At any time t, what is known about a particular pixel is its intensity history values. A recent history of each pixel $\{P_1,.........,P_t\}$ is modeled by a mixture of K Gaussian distributions. The probability of observing the current pixel value (C_t) is:

$$P(C_t) = \sum_{i=1}^{K} \omega_{i,t} * \eta(X_t, \mu_{i,t}, \sigma i, t), \tag{6}$$

where

K: the number of Gaussian functions to model the pixel

$\omega_{i,t}$:the estimated weight of the ith Gaussian in the mixture at time t

$\mu_{i,t}$: the mean value of the ith Gaussian in the mixture at time t

$\sigma i, t$: the standard deviation of the ith Gaussian in the mixture at time t, and

$$\eta(x_t,\mu,\sigma) = \frac{1}{(2\pi)^{n/2}|\sigma|^{1/2}} e^{\frac{-1}{2}(x_t-\mu_t)^T \sigma^{-1}(x_t-\mu_t)}, \tag{7}$$

Every new pixel value, P_t, is checked against the existing K Gaussian distributions until a match is found. A match is defined as a pixel value within 2.5 of the standard deviation σ of a distribution.

The maintenance of the model is done with a new pixel based on the pixel to GMM match. There are two cases for the maintenance as following:

Case one: If none of the K distributions match the current pixel value, then the least probable distribution is replaced by a distribution with the current pixel value as its mean value, an initially high variance, and low prior weight

$$\mu_t = P_t \tag{8}$$

$$\omega_{i,t} = \alpha \tag{9}$$

where α is the learning rate of the GMM

Case two: If one or more distribution functions match the new pixel value, then the matched functions' parameters are updated as following:

$$\mu_t = (1-\rho)\mu_{t-1} + \rho P_t \tag{10}$$

$$\sigma_t^2 = (1-\rho)\sigma_{t-1}^2 + \rho(P_t-\mu_t)^T(P_t-\mu_t) \tag{11}$$

where

$$\rho = \alpha\eta(P_t|\mu_k,\sigma_k) \tag{12}$$

The weights of the K distribution are updated as following:

$$\omega_{t,k} = (1-\alpha)\omega_{k,t-1} + \alpha M_{k,t} \tag{13}$$

where $M_{k,t}$ is 1 for models that are matched and 0 for the remaining distributions. The $M_{k,t}$ and σ parameters for the unmatched distributions remain the same.

To estimate the foreground pixels of the current frame using the GMM model, the Gaussian distributions for each pixel are sorted in descending order based on the value ω/σ. This value increases as a distribution gains more evidence to represent a background pixel. The most likely background distributions remain on top and the less probable transient background distributions gravitate towards the bottom and are eventually replaced by new distributions. The algorithm selects the first B distributions that counts for a predefined fraction of the evidence T.

$$B = argmin_b\{\sum_{i=1}^{b}\omega_i > T\} \tag{14}$$

where

B: the distributions that represent the background model

ω_i: the weight of the distribution i

T: a threshold for the minimum background ratio to the image, usually 0.7.

The output of the GMM is a digital image with values of zero or one. Ones represents the foreground pixel and appears in white color. Zeroes represents the background pixels and appears in black color. GMM shows a very good result for foreground extraction when there are statically moving objects, such as moving trees due to a wind. It also shows a good maintenance for the background with removed and inserted background objects. The output image of the GMM is called foreground digital mask.

Fig. 4 shows an example compares between the mean filter and the GMM in foreground pixels extraction. 70 images were captured for a road intersection at different time stamps and used for the background modeling, the time separation between the frames is 10 sec. The left image shows a vehicle that was inserted in the background in the last 30 image frames, this car shall be categorized as background object. The mean filter has detected the car as foreground, while GMM adapted to the inserted object (the car) quickly and categorized it as background.

C. Morphological Filtering

The background model using GMM may have false positives in some regions of the image due to statically moving objects, and objects that were removed or added to the background. It can also miss-detect foreground pixels due to the similarity of the foreground pixels and the background. Morphological filtering removes the noise in the foreground digital mask by connecting the neighbor foreground regions to construct the shape of the objects. It disconnects the small and the outlier foreground regions that doesn't belong to the same object. It also closes the small holes in the foreground digital mask.

Morphological image processing is suitable for binary image processing since it depends only on the relative ordering of pixel values, and not on their numerical values. Morphological operations are a collection of non-linear operations related to the shape or morphology of features in an image. More details about morphological filtering can be found in [21].

There are two fundamental operations for morphological filtering, erosion and dilation. Also, there are compound operations by mixing the erosion and the dilation. Opening filter is erosion followed by dilation. closing filter is a dilation followed by an erosion. Fig. 5 shows an example of a binary image filtered with opening filter and closing filter. As shown in the in the figure, the opening filter (the center image) connects the close foreground regions together, which helps constructing the shape pf the foreground object. The closing filter (the right image)removes the small foreground regions, which helps in removing the small false foreground extractions. After trying many morphological filters with many sizes and structures, observations showed that filtering an image with a 10x10 square closing filter followed by a 3x3 opening filter provides the best result to remove the noise from the foreground digital mask and connect the foreground regions. The closing filter constructs the shape of the moving regions by connecting them together. The opening filter removes the small holes in the image. Fig. 6 shows examples of the foreground digital mask filtering using a square closing filter with size of 10x10 followed by an opening filter with size of 3x3.

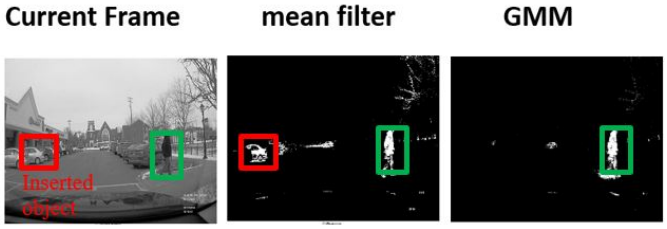


Fig. 4. The first image shows the image frames, the second image shows the foreground digital mask using the mean filter, and the right image shows the foreground digital mask using GMM. The true positives are highlighted in green, while the false positives are highlighted in red

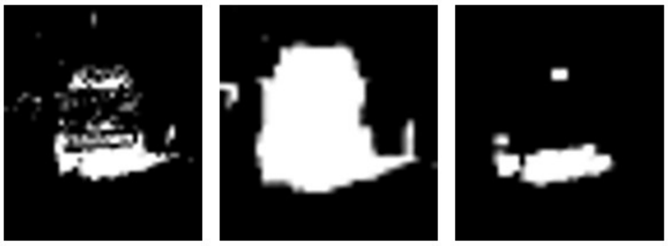


Fig. 5. The first image is the foreground digital mask of a moving object using GMM, the second image shows the output image after applying a closing filter, and the third image is the output of implementing the opening filter

D. Foreground Digital Mask Thresholding for Candidate Generation

A typical candidate generation scans the whole image or a large portion of it to generate the candidates. In our infrastructure system, the generation of the candidates focuses on the foreground regions only and excludes the background objects.

The candidate generation in the infrastructure algorithm applies a threshold to the foreground digital mask. The digital mask is scanned by a sliding window of a 64x128. The mean of the window is calculated; if the mean is higher than the threshold, the same window in the corresponding image is passed to the next step, and, if not, the region is excluded from being a candidate and the scanning window moves to the next region in the digital mask.

Fig. 7 shows the flowchart of the candidate generation. The image is scanned at multiple sizes of its original size. The scanning is applied on the images while resolution is varied from 0.5 to 1.3 with a step of 0.1. Fig. 8 shows an example of how the candidate generation approach is applied on an image. The main advantage of the candidate generation using infrastructure background modeling is to reduce the number of the candidates from the static regions. This reduction is reflected in the performance of the detection algorithm as shown in the system evaluation section.

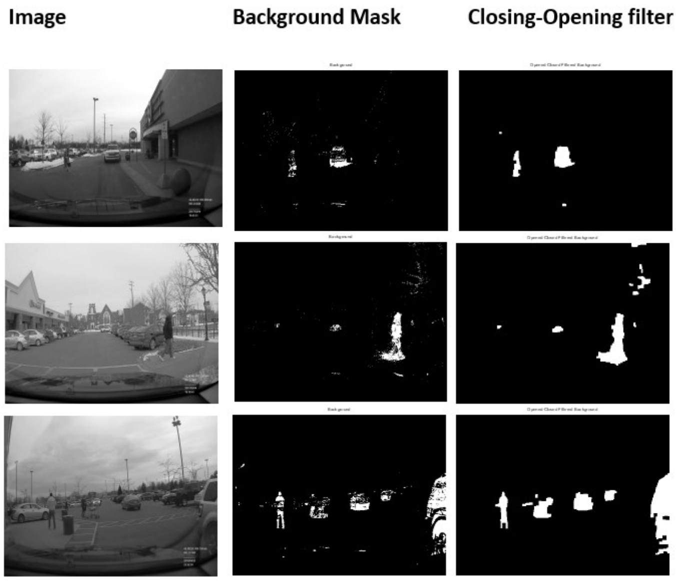


Fig. 6. The first column shows three input images, the second column is the foreground digital mask using GMM, and the third column is the filtered mask using a square closing filter of 10x10 followed by a square opening filter of 3x3

E. HOG and SVM for Feature Extraction and Candidate Generation

Pedestrians are one of the most complex objects to detect because they can appear in different sizes, poses, and colors. The shape of the pedestrian may change while carrying different objects. The change in the outdoor light conditions is another challenge. To go over these challenges, unique features of the object are extracted to provide a robust description of pedestrians. These features can be textures, contours, and edges. The features of the object should be very similar under different view conditions.

Histogram Oriented Gradient (HOG) feature extraction is considered as one of the most successful approaches for pedestrian detection when it is used with Support Vector Machine (SVM) classifier. This model was introduced by Dalal and Triggs in 2005 [22]. The main advantages of

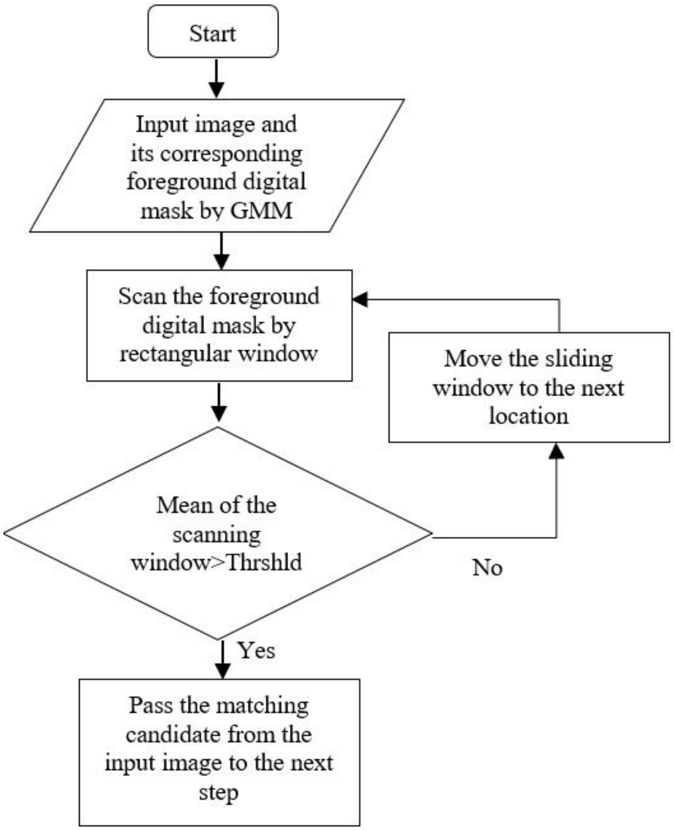

Start
Input image and its corresponding foreground digital mask by GMM
Scan the foreground digital mask by rectangular window
Move the sliding window to the next location
Mean of the scanning window>Thrshld
No
Yes
Pass the matching candidate from the input image to the next step

Fig. 7. Candidate generation flow chart

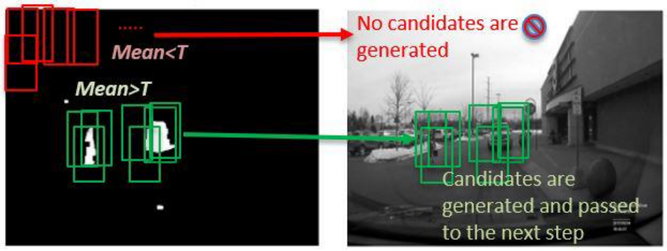

Mean<T
Mean>T
No candidates are generated
Candidates are generated and passed to the next step

Fig. 8. The left image shows the candidate generation, the right image comparing the mean of the scanning window to a threshold to make the decision for candidate generation

HOG are the induced robustness against the global and the local illumination changes, the moderation of pedestrian pose differences, and algorithm runtime. HOG with SVM are used for pedestrian detection in our infrastructure module due to its accurate detection result.

HOG is calculated by computing the first order image gradients. It captures the object contours and the texture information. Features are collected in a vector and passed to the classifier. Dalal and Triggs explained the mathematical model of the algorithm and analyzed the detection results using many human datasets [22].

SVM is a learning model that analyzes the training data and build a set of rules to classify similar observations that haven't been seen before. SVM requires training data for each class. In our case, the classes are pedestrian and non-pedestrian. HOG vectors are passed to the SVM to develop the classification rules. More information about the SVM model can be found in [23]. One of the main advantages of the SVM is the ability to use Kernel functions to transfer the data to a higher dimensional domain to provide an accurate classification for the non-linearly separable data. More information about the training data and the used SVM parameters are listed in the implementation section. Fig. 9 shows the block diagram for the HOG with the SVM for pedestrian detection. Fig. 10 shows the block diagram of the infrastructure pedestrian detection algorithms including the image background modeling and moving object detection, image filtering, image thresholding and candidate generation, HOG, and SVM for pedestrian classification.

IV. The Infrastructure Pedestrian Detection System Implementation

A. Testing and Training Datasets

To implement the infrastructure system, a labeled dataset is required for SVM. A test dataset is also needed to verify the system results. There are many pedestrian datasets available online, such as INRIA and MIT. However, none of these datasets can be used to implement the infrastructure system since multiple images for the same location at different time stamps are required for background modeling.

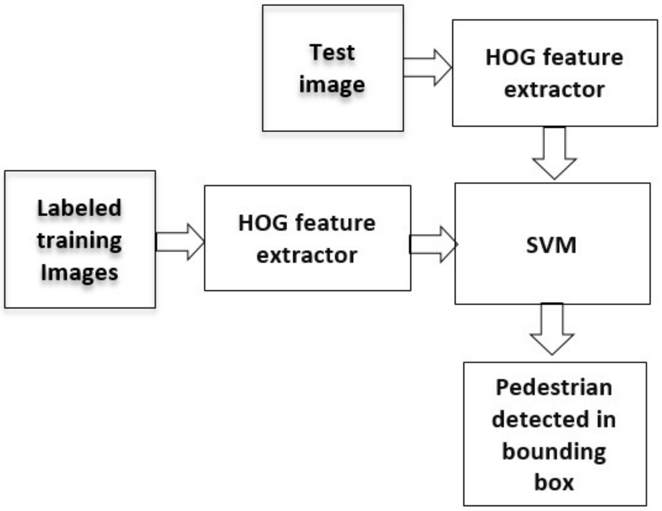


Fig. 9. Histogram Oriented Gradient (HOG) and Support Vector Machine (SVM) for pedestrian detection block diagram

A vehicle was setup with a front windshield camera for video collection. The camera is equipped with external memory to store the videos. The used camera is a 3.2 MP CMOS sensor with a 135 degrees field of view. The output video resolution is1280x720 with 60 fps.

1) The training dataset: Videos were collected from locations with pedestrian traffic, such as downtowns, shopping centers, and school campuses. Videos were sampled to frames, and the "Training Image Labeler" MATLAB tool was used to label pedestrians in the images. The training data include 1066 positive samples and 1600 negative samples. Fig. 11 shows an example of positive samples for pedestrians, and negative samples like trees and buildings.

2) The testing dataset: Testing videos were collected while the vehicle was stationary to capture multiple images for the

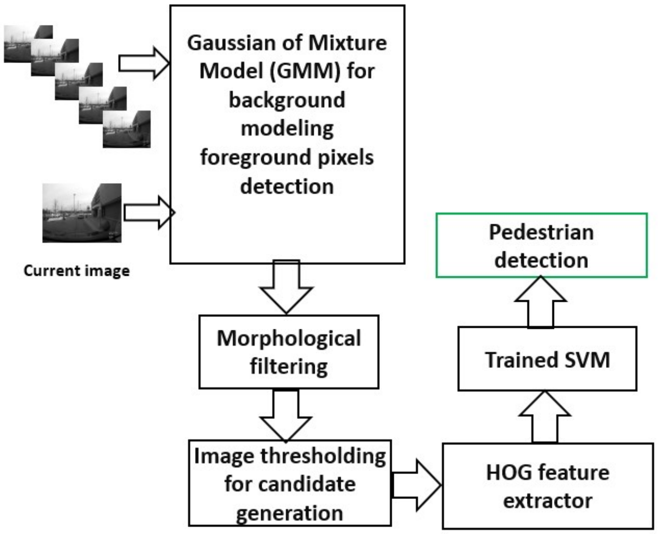

Gaussian of Mixture Model (GMM) for background modeling foreground pixels detection
Current image
Morphological filtering
Image thresholding for candidate generation
HOG feature extractor
Trained SVM
Pedestrian detection

Fig. 10. Infrastructure image processing module algorithms for pedestrian detection

same location. The testing data were collected from different locations with vehicle and pedestrian traffic. By sampling the collected videos for each location to frames, many image frames for the same location at different times are available for background modeling. The test dataset specifications are as following:

- The testing data collected for 100 different locations
- The video length for each location is 330 sec
- Each video is sampled to frames with rate of 1 fps
- Pedestrians are labeled in the last image frame for each location

The last image frame represents the current image frame of the passing vehicle, where the algorithm shall detect the pedestrians in it. The separation time between the frames can be selected by the algorithm. For example, if the algorithm reads one frame every 10 seconds, this represents a vehicle passing from the location every 10 seconds. The separation time between the frames represents the road traffic. The time separation impact in the detection is studied later in this section. The first frame represents the reference frame for image registration. Fig. 12 shows an example of testing image frames for a location; it shows the previous frames and the current image frame with the labeled pedestrian.

B. MATLAB Implementation

The infrastructure pedestrian detection system is implemented using MATLAB. The implementation is divided in blocks as following:

- Testing image frames read:

  In this block, the image frames for a location is imported to MATLAB. The time separation between the frames (T sep) can be selected by the algorithm to simulate the different traffic conditions. This time represents the time between the passing vehicles.
- Image registration:
  Images for each location were registered to the reference frame using Harris-Stephens approach as explained above.

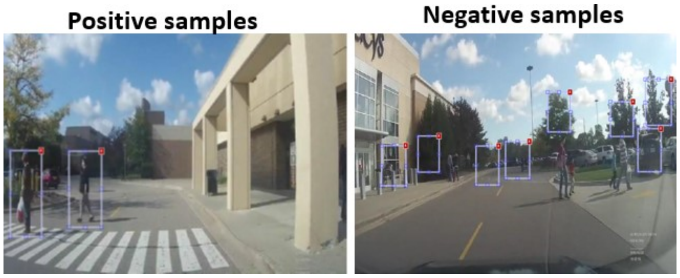


Fig. 11. The left image shows labeled positive samples and the right image shows labeled negative samples. The images were labeled using the MATLAB toolbox “Training Image Labeler”

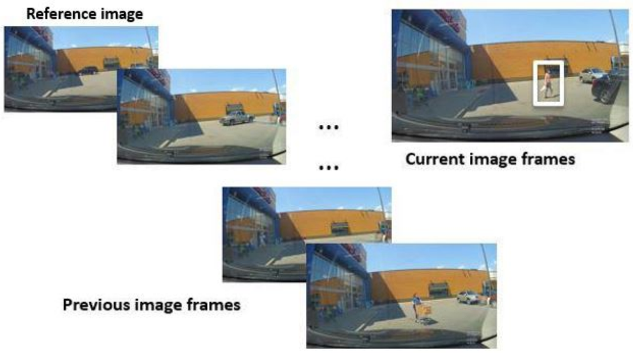


Fig. 12. Testing images for a location, it includes the previous image frames and the current image with a labeled pedestrian

- GMM for foreground extraction:
  The imported images for a location are passed to the GMM for background modeling and foreground pixels extraction. Table I summarizes the GMM parameters and their nominal values used in the implementation.
- Morphological filtering:

  The foreground digital mask is filtered using a 10x10 square closing filter to connect the foreground regions followed by a 3x3 square opening filter to close the small holes in the foreground mask.
- Candidate generation:

  The foreground digital mask is scanned by a 64x128 window at multiple sizes from 0.5 to 1.3 with a step of 0.1. If the mean of the window is greater than a threshold, the candidate is passed to HOG. The threshold value impact in the detection result is studied in the next section.
- HOG with SVM:

  HOG is used to extract the features of the candidates. Each candidate produces 3780 features. SVM is trained using 1066 positive samples and 1600 negative samples. The implemented HOG main parameters are shown in Table II.

TABLE I. THE NOMINAL VALUES FOR GMM PARAMETERS IN THE MATLAB IMPLEMENTATION

GMM parameters	Nominal values
Learning rate (α)	0.005
Maximum background ratio	0.7
Initial variance	900
Number of Gaussians	5

C. Frames Separation Time Impact on the Detection (T_{sep})

The time separation between the frames is an important factor that affects the system, as it specifies in which traffic situations the system can be implemented. Urban areas such as downtowns and shopping centers usually have a high vehicle traffic, so the time separation is short, while it is longer in rural areas with low traffic.

The time separation factor (T_{sep}) is studied for the following values: 5 seconds, 10 seconds, 20 seconds, 30 seconds, and 40 seconds. Fig. 13 shows the Receiver Operating Characteristic (ROC) of the precision and the recall for each time separation value; the performance of the system is very similar for all the (T_{sep}).

The result shows that the system provides a good detection result under many traffic conditions. However, in low traffic conditions, there is more chance to miss some changes in the background. This will result in false foreground detection, which means generating more candidates from background regions.

V. Infrastructure System Evaluation

For better understanding of the infrastructure pedestrian detection system results, a comparison of the detection results is done with a reference approach of a traditional on-vehicle pedestrian detection system. Detection results were compared with respect to the detection accuracy by counting false positives and false negatives reported by each system. The runtime of the reference algorithm was also compared to the infrastructure system to show the effect of the improved candidate generation of the infrastructure system in the processing time of the detection.

TABLE II. The nominal values for the HOG parameters in the MATLAB implementation

HOG parameters	Nominal values
Cell size	8x8 pixels
Block size	2x2 cells
Block overlapping	50%
Number of histogram bins	9

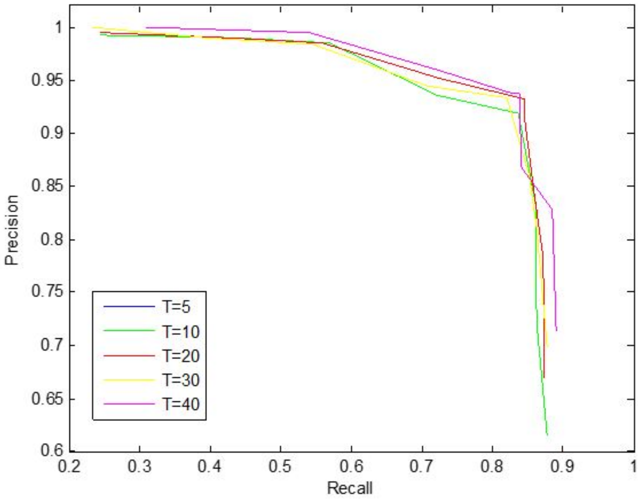


Fig. 13. Precision vs. Recall ROC plot for different frame separation time (T_{sep})

In the reference algorithm, the candidates were generated using multiple size image scanning. The input image is scanned with a fixed scanning window of 64x128. The scanning included the whole image except the top of the images that includes the sky. Images were scanned between 0.5 to 1.3 of their size, with a step of 0.1.

Candidates were passed to HOG for feature extraction, and then a trained SVM was used for classification. Fig. 14 shows the block diagram of the reference pedestrian detection system and the proposed infrastructure pedestrian detection system. The blocks colored in green are common between the reference and the infrastructure system. The blue blocks are related to the on-vehicle detection system, while the yellow ones are related to the proposed infrastructure system. That means any improvement in the detection result in the infrastructure system is related to the improved candidate generation using background modeling and foreground pixels extraction.

A. Detection Results

The testing dataset for the 100 locations were passed to the infrastructure system for pedestrian detection. The labeled testing frames were also passed to the reference algorithm. No previous images were used in the reference algorithm since there is no background modeling.

The number of the generated candidates by the reference algorithm using multiple size image scanning was 42900 for the whole dataset. The total number of the generated candidates using the infrastructure algorithm was 28750. The first advantage of the infrastructure system is the reduction in the number of the candidates by 33% when compared to the reference algorithm.

The infrastructure system reported 24 false positives. The reference algorithm reported 98 false positives. This shows a 75.5% reduction in false positives in the infrastructure system. This significant improvement is due to the reduction in the number of candidates that is generated from the background region that may cause more false positives in the reference algorithm.

The infrastructure system showed better results in false negatives compared to the reference algorithm. The total number of false negatives reported by the infrastructure model is 19, as compared to 24 for the reference algorithm. The reason for the reduction in the false negatives in the infrastructure is that candidate generation is focused in the foreground pixels, which increases the possibility of capturing candidates for a pedestrian in different poses and angles, thereby increasing

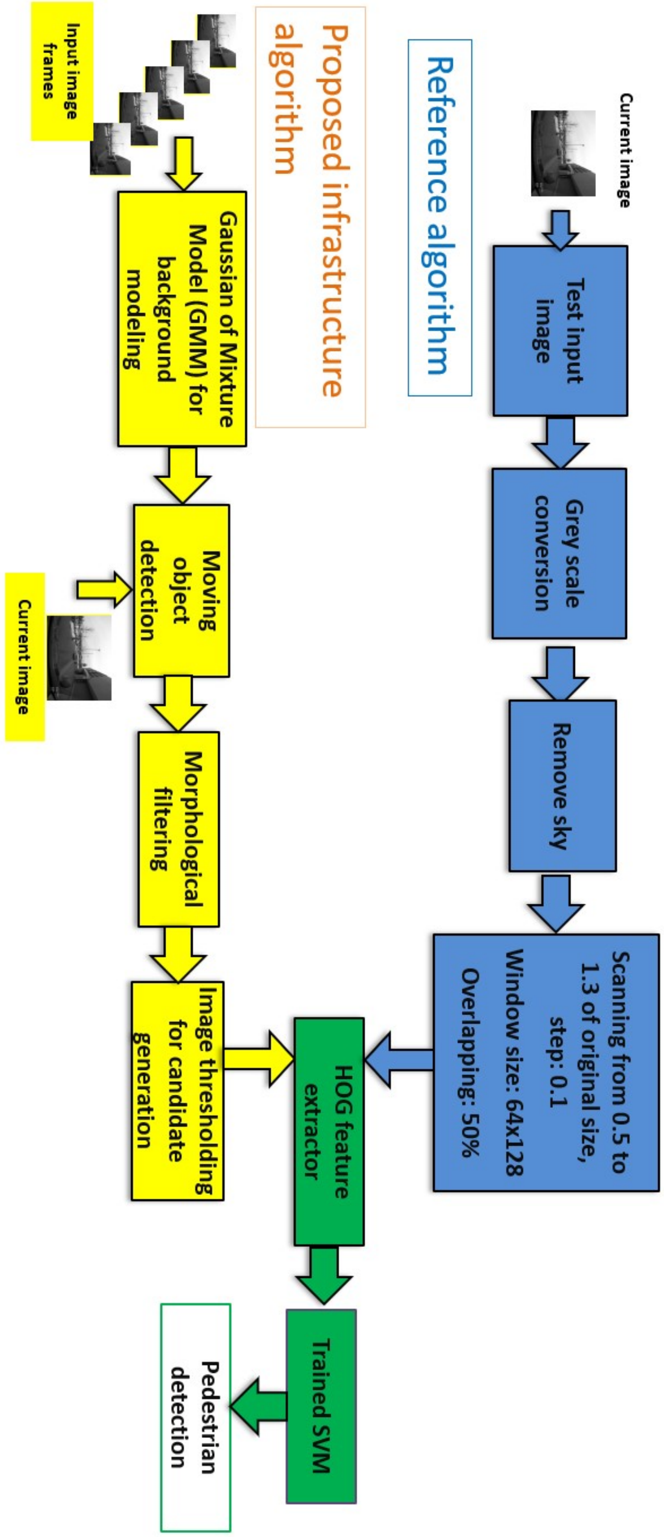


Fig. 14. The infrastructure Pedestrian Detection system and the reference algorithm block diagram

the chance to classify the candidates correctly. Table III summarizes the detection results of the infrastructure system and the reference system.

Fig. 15 shows the Receiver Operating Characteristic (ROC) curve of the precision and the recall for the infrastructure system and the reference system. The infrastructure system shows very high precision values when compared to the reference algorithm. It also shows a better recall at many SVM operating points.

Fig. 16 shows an example of the reference algorithm detection compared to the infrastructure algorithm. The reference algorithm showed a false positive for a background object highlighted in red, while the infrastructure system didn't report the same false positive. Fig. 17 shows another example of a pedestrian miss-detection in the reference algorithm, while it is detected in the infrastructure algorithm.

B. Algorithm Runtime

One of the main advantages of the infrastructure algorithm is the reduction in the number of the candidates, which reduces the runtime of the detection system. The runtime of the infrastructure system was compared to the reference detection system by computing the time to process and classify the testing dataset. The computer specifications used for the runtime study are listed below:

- Processor: Intel(R) Core(TM) i7-8550U CPU @ 1.8

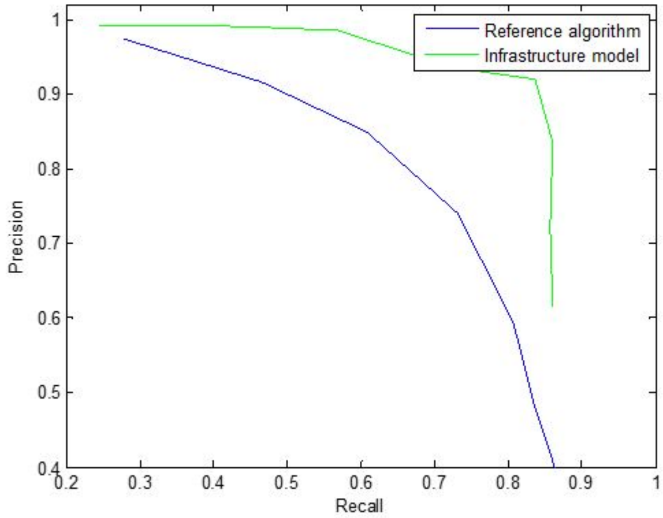


Fig. 15. Precision vs. Recall ROC for the infrastructure system and the reference system

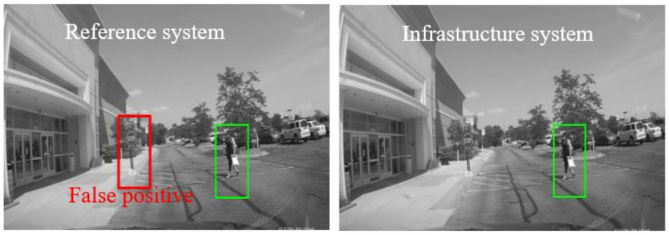


Fig. 16. Left image shows a false positive for the stop sign reported by the reference algorithm, the right image shows the detection result for the same image with no false positive using the infrastructure model

TABLE III. The detection results summary for the infrastructure model and the reference algorithm

Method	Total candidates	False negative	False positive	Recall	Precision
Infrastructure system	28750	19	24	0.837	0.894
The reference system	42900	24	98	0.807	0.593

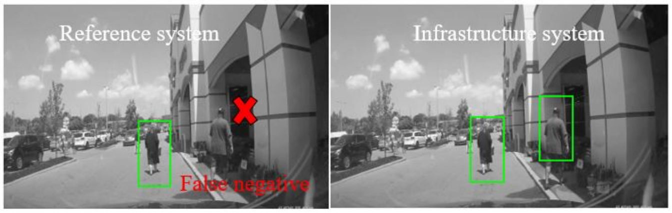


Fig. 17. The left image shows pedestrian miss-detection using the reference algorithm, the infrastructure model detected the pedestrian in the same image

GHz

- RAM: 12 GB

- System type: 64-bit operating system

The runtime of the reference system for an image frame is given by:

$$RunTime_{ref} = Number of candidates * T_{HOG/SVM} \quad (15)$$

where
$RunTime_{ref}$ is the runtime for one image frame using the reference algorithm measured in sec/frame.

$T_{HOG/SVM}$ is the reference algorithm runtime to classify one candidate in sec/candidate.

$T_{HOG/SVM}$ equals 0.0558 sec/candidate. the number of candidates generated by the reference algorithm is 429 per frame. By applying Equation (13), the runtime for the reference algorithm is 23.938 sec / frame.

The runtime for the infrastructure system for an image is given by:

$$RunTime_{inf} = T_{GMM} + \\ Number of candidates * T_{HOG/SVM} \quad (16)$$

where

$Runtime_{infrastructure}$ is the runtime of the infrastructure algorithm for one candidate measured in sec/frame

T_{GMM} is the time to extract the foreground pixels from the current image frame and maintain the background model, the time is measured in sec/frame.

$T_{HOG/SVM}$ is the reference algorithm runtime to classify one candidate in sec/candidate.

$T_{HOG/SVM}$ pf the infrastructure system has the same value in the reference algorithm because HOG and SVM are common processes in the two approaches.

T_{GMM} equals 0.0484 sec/frame in both approaches. The total number of candidates in the infrastructure algorithm is reduced by 33% when compared to the reference algorithm. Therefore, the total number of candidates per image frame using the infrastructure algorithm equals to 287.43 candidate/frame.

By applying Equation (14), the runtime for the infrastructure algorithm equals 16.086 sec/frame. The analysis shows that the runtime of the infrastructure algorithm is reduced by 32.7% when compared to the reference algorithm. Table IV summarizes the runtime analysis for the infrastructure algorithm and the reference algorithm.

VI. Conclusions

This paper proposed a system to improve the candidate generation process for pedestrian detection in connected vehicles. The system registers the collected images for a location to a reference image. Harris-Stephens approach for corner detection, Nearest Neighbor Distance Ratio (NNDR) for feature mapping and image transformation are used in the registration step.

Gaussian Mixture Model (GMM) is used to model the background of a location using the registered images stored in the infrastructure database. The foreground pixels in the images extracted using the GMM model. Candidates are generated through scanning the foreground regions by a rectangular box. Finally, Histogram Oriented Gradient (HOG) and Support Vector Machine (SVM) were used to classify candidates as pedestrians or non-pedestrians.

A data-set is collected for algorithm training and test. A reference algorithm is implemented to highlight the

TABLE IV. The runtime for the infrastructure model and the reference algorithm

Detection system	Algorithm runtime (sec/frame)
Infrastructure system	16.086
Reference system	23.938

improvements achieved in the proposed system.

The infrastructure pedestrian detection system showed a huge improvement in detection performance when compared to the reference algorithm that represents a typical on-board detection approach. The infrastructure algorithm significantly reduced the number of the generated candidates when compared to the reference algorithm. The generated candidates in the proposed infrastructure system is reduced by 33%. Also, the false positives are reduced by 75% in the infrastructure system compared to the reference algorithm. Since the infrastructure system classifies less candidates, the runtime of the algorithm is improved by 67% when compared to the reference algorithm.

REFERENCES

[1] World Health Organization. Global status report on road safety 2015. World Health Organization, 2015.

[2] Papageorgiou, Constantine, and Tomaso Poggio. ”A trainable system for object detection.” International journal of computer vision 38, no. 1 (2000): 15-33.

[3] Broggi, Alberto, Massimo Bertozzi, Alessandra Fascioli, and Massimiliano Sechi. ”Shape-based pedestrian detection.” In Intelligent Vehicles Symposium, 2000. IV 2000. Proceedings of the IEEE, pp. 215-220. IEEE, 2000.

[4] Broggi, Alberto, Paolo Grisleri, Thorsten Graf, and M. Meinecke. ”A software video stabilization system for automotive oriented applications.” In Vehicular Technology Conference, 2005. VTC 2005-Spring. 2005 IEEE 61st, vol. 5, pp. 2760-2764. IEEE, 2005.

[5] Bombini, Luca, Pietro Cerri, Paolo Grisleri, Simone Scaffardi, and Paolo Zani. ”An evaluation of monocular image stabilization algorithms for automotive applications.” Intel. Transp. Syst (2006).

[6] Llorca, D. F., M. A. Sotelo, A. M. Hellín, A. Orellana, M. Gavilán, I. G. Daza, and A. G. Lorente. ”Stereo regions-of-interest selection for pedestrian protection: A survey.” Transportation research part C: emerging technologies 25 (2012): 226-237.

[7] Najm, Wassim G., Jonathan Koopmann, John D. Smith, and John Brewer. Frequency of target crashes for intellidrive safety systems. No. DOT HS 811 381. United States. National Highway Traffic Safety Administration, 2010.

[8] Belyaev, Evgeny, Alexey Vinel, Adam Surak, Moncef Gabbouj, Magnus Jonsson, and Karen Egiazarian. ”Robust vehicle-to-infrastructure video transmission for road surveillance applications.” IEEE Transactions on Vehicular Technology 64, no. 7 (2015): 2991-3003.

[9] Pervez, Farhan, Abdulkareem Adinoyi, and Halim Yanikomeroglu. ”Efficient resource allocation for video streaming for 5G network-to-vehicle communications.” In Personal, Indoor, and Mobile Radio Communications (PIMRC), 2017 IEEE 28th Annual International Symposium on, pp. 1-6. IEEE, 2017.

[10] Zitova, Barbara, and Jan Flusser. ”Image registration methods: a survey.” Image and vision computing 21, no. 11 (2003): 977-1000.

[11] Gerónimo, David, and Antonio M. López. Vision-based pedestrian protection systems for intelligent vehicles. New York, NY, USA:: Springer, 2014.

[12] Al-Refai, Ghaith, Modar Horani, and Osamah Rawashdeh. A Framework for Background Modeling Using Vehicle-to-Infrastructure Communication for Improved Candidate Generation in Pedestrian Detection. 17th Annual IEEE International Conference on Electro Information Technology EIT, 2018.

[13] Harris, Chris, and Mike Stephens. ”A combined corner and edge detector.” In Alvey vision conference, vol. 15, no. 50, pp. 10-5244. 1988.

[14] Li, Qiaoliang, Guoyou Wang, Jianguo Liu, and Shaobo Chen. ”Robust scale-invariant feature matching for remote sensing image registration.” IEEE Geoscience and Remote Sensing Letters 6, no. 2 (2009): 287-291.

[15] 14. Lee, B., and M. Hedley. ”Background estimation for video surveillance.” (2002).

[16] McFarlane, Nigel JB, and C. Paddy Schofield. ”Segmentation and tracking of piglets in images.” Machine vision and applications 8, no. 3 (1995): 187-193.

[17] Wren, Christopher Richard, et al. ”Pfinder: Real-time tracking of the human body.” IEEE Transactions on pattern analysis and machine intelligence 19.7 (1997): 780-785.

[18] Stauffer, Chris, and W. Eric L. Grimson. ”Adaptive background mixture models for real-time tracking.” In cvpr, p. 2246. IEEE, 1999.

[19] Bouwmans, Thierry. ”Traditional and recent approaches in background modeling for foreground detection: An overview.” Computer Science Review 11 (2014): 31-66.

[20] Bouwmans, Thierry. ”Recent advanced statistical background modeling for foreground detection-a systematic survey.” Recent Patents on Computer Science 4, no. 3 (2011): 147-176.

[21] Serra, Jean. ”Morphological filtering: an overview.” Signal processing 38, no. 1 (1994): 3-11.]

[22] Dalal, Navneet, and Bill Triggs. ”Histograms of oriented gradients for human detection.” In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1, pp. 886-893. IEEE, 2005.

[23] Burges, Christopher JC. ”A tutorial on support vector machines for pattern recognition.” Data mining and knowledge discovery 2, no. 2 (1998): 121-167.

Intelligent Parallel Mixed Method Approach for Characterising Viral YouTube Videos in Saudi Arabia

Abdullah Alshanqiti[1]
Faculty of Computer and Information Systems
Islamic University (IU)
Madinah, Saudi Arabia
ORCID:0000-0002-6080-5236

Ayman Bajnaid[2]
Faculty of Media and Communication
King Abdulaziz University (KAU)
Jeddah, Saudi Arabia

Abdul Rehman Gilal[3]
Department of Computer Science
Sukkur IBA University
Sukkur, Pakistan

Shuaa Aljasir[4]
Faculty of Media and Communication
Kin6g Abdulaziz University (KAU)
Jeddah, Saudi Arabia

Aeshah Alsughayyir[5]
College of Computer Science
and Engineering
Taibah University
Madinah, Saudi Arabia

Sami Albouq[6]
Faculty of Computer
and Information Systems
Islamic University (IU)
Madinah, Saudi Arabia

Abstract*—In social networking platforms, comprehending virality, exemplified by YouTube, is of great importance, which helps in understanding what characteristics utilised to create content along with what dynamics involved in contributing to YouTube's strength as a platform for sharing content. The current literature surrounding virality problem appears sparse concerning development theories, investigations regarding empirical facts, and an understanding of what makes videos go viral. The overarching objective is to understand deeply the phenomena of viral YouTube videos in Saudi Arabia, hence we propose an intelligent convergent parallel mixed-methods approach that begins, as an internal step, by a qualitative thematic analyses method and an NLP-based quantitative method independently, followed by training an unsupervised clustering model for integrating the internal analysis outputs for deeper insights. We have empirically analysed some trended YouTube videos along with their contents, for studying such phenomena. One of our main findings revealed that boosting entertainments, traditions, politics, and/or religion issues when making a video, that is associated in somehow with sarcastic or rude remarks, is likely the preeminent impulse for letting a regular video go viral.

***Keywords*—Virality; text mining; sentiment analysis; social media analysis; mixed method approach**

I. Introduction

Knowing how digital content can get rapidly spread worldwide, such as viral video, is of great importance in perfecting our e-services. In the scope of social networking platforms, virality can be loosely defined as the ability of content to spread rapidly in society from one person to another. Given the present time's propensity for communication via electronics, content spreads like wildfire thanks to the Internet. This virality is exemplified by YouTube, whose user-generated content allows users to freely create and share content both on its own platform and on other social media platforms [1].

Given YouTube's success, there exists an interest in understanding what characteristics utilised to create content along with what dynamics involved in contributing to YouTube's strength as a platform for sharing content. Although scholars agree with the characteristics that constitute/make up viral content, there exists less certainty with what makes a video extremely popular [2]. Despite a growing interest in this field, the current literature surrounding this topic also remains sparse with respect to development theories, investigations regarding empirical facts, and an understanding of what makes videos go viral.

Understanding why and how a video becomes extremely popular (i.e., how it goes viral) can maximise how consumers can benefit from a video's popularity along with how users can deal with the threats associated with virality such as spreading rumors or violating others' privacy. Analysing a large amount of data from YouTube's video collection would also allow for a deeper understanding of social behavior, dynamics, and processes at play when people consume and create content.

Broadly speaking, there exists two principal conceptual analysis when it comes to research on virality, formulated coherently in a valuable theoretical framework by [3]: a top-down mechanism which considers virality as the result of highly influential individuals who can use their power in promoting their videos by (e.g., existing mainstream media); a bottom-up mechanism, which argues that virality relies instead on the characteristics of the content that factually engage individuals to spread the content in a self-motivated way [4]. Interestingly, [5] (citied in [6]) mention that the latter mechanism is more often prompting virality.

In a general sense, this research attempts to contribute to the bottom-up mechanism by solely focusing on Arabic videos, particularly, videos that have gone viral. The overarching objective behind our attempt is to provide an intelligent based solution to help in understanding deeply the phenomena of viral YouTube videos in Saudi Arabia, which can be used in future research as a guideline or for comparison purposes. Thus, we propose a convergent parallel mixed-methods approach that begins, as an internal step, by a qualitative thematic analyses method and an NLP-based quantitative method independently, followed by training an unsupervised clustering model for integrating the internal analysis outputs for deeper insights. To be more precise, the proposed complex approach depends on (1) our optimised lexicon-based *Bag of Words* sentiment classifier for analysing viewer's shared comments left on YouTube, and on (2) a manual inspiration method for qualitatively analysing video content. We report on experiments to understand the virality problem by examining several trended YouTube videos in Saudi Arabia. Summing up, the contributions of this research are:

- A qualitative study on a variety of video's categories and themes propagated in Saudi Arabia.
- A lexicon-based *Bag of Words* sentiment classifier, where the novelty here lies on our optimised algorithms, implemented in Java, that support any texts written in Arabic without translation.
- An innovative idea of utilising unsupervised machine learning technique, depending on distance matrix and hierarchical clustering, for integrating our internal findings. This thought could be a promising research paradigm that fundamentally contributes to social media intelligence approaches.

The next section seeks to investigate prior scholarship on phenomena that have gone viral, examine gaps in the literature regarding the virality process, and present noteworthy questions of the current research. The following section introduces our methodology utilised in the examination and presents the subsequent analysis and results. Lastly, the final section discusses the conclusions of this research and outline our intention for future research to take.

II. Review of Related Literature

This section provides an overview of the phenomenon of viral content by drawing on scholars who have sought to understand the processes and dynamics of virality. In 1997, the firm Draper Fisher Jurvetson coined the phrase “viral marketing” to describe Hotmail’s use of advertisements to promote the fact that its emailing service was free [7]. [8] then noted that viral marketing was described as a type of marketing that infects its customers with an advertising message that passes from one customer to the next like a rampant flu virus (p. 93). More generally, “viral marketing” and “viral content” have since become catchphrases for online advertising success. A variety of other definitions have also been offered for virality, each coupled with a specific approach in examining its nature.

According to [9], examinations and definitions of virality can be categorised in three ways. The first seeks to examine how the content is accessed, disseminated, and propagated over a short time period. The second seeks to examine how virality is spread via electronic sharing (i.e., word of mouth) by focusing on the content shared. Lastly, the third concentrates on users’ behaviors and engagement with the viral content in question and gauges their likes/dislikes, shares, and comments. [10] argued that the term “virality” includes a host of aspects and exchanges such as the number of people who have access to the content, the appreciation of the content, and how many people have liked or shared the content. The popularity of the content depends exclusively on those who share it and the reactions it garners (positive, negative, and, to a lesser extent, neutral). The current research defines viral content as that which spreads to the greatest degree possible over the shortest amount of time.

YouTube has been chosen as the topic of study for the present research due to the double-sided nature of its platform (i.e., the ability to share and participate through comments as well as to react to content via word of mouth). Sharing content on YouTube requires interacting with others online, which in turn affects the popularity of said content. Content spread online generates greater audience numbers than content spread through some other means. YouTube also affords the distinct opportunity to study both the activities of YouTube users’ interactions and their social network ties. According to [10], a number of elements play a part in this sharing process, including the nature of the shared content, the nature of the user who shares it, the nature of the audience who receives it, and the structure of the network through which the content is spreading. The present research aimed to provide an AI-based solution to help in understanding the phenomena of viral YouTube videos.

Previous research on virality has primarily been drawn from five different fields: psychology (e.g., [9]; [11]; [12]; [13]; [4]), computer science (e.g., [14]; [15]; [16]; [1]; [10]; [17]; [18]; [19]; [20]; [21]; [22]), political science (e.g., [2]; [23]; [24]), marketing (e.g., [25]; [26]; [27]; [28]; [29]; [30]; [31]; [7]; [32]; [33]), and health (e.g., [34]). These studies have been mainly conducted in Western countries, such as the United States, Canada, Germany, Italy, and Australia. However, there have been a few studies conducted in less developed countries, such as [2] study in South Africa; [29] and [22] studies in China; [1] in South Korea; [18] in Brazil; [4] in Romania; and [21] in India. However, no studies have been conducted in Arab countries or even in the Middle East.

These studies used several methods to collect data. While some of them utilised questionnaires to obtain users responses, others used data-mining tools. A few studies manually conducted content analysis. Most of the previous researchers developed their own models to explain the phenomenon of virality. Only two studies have borrowed theories from other fields to explain virality; these theories included uses and gratifications theory, the persuasion model, and the memory-based model ([9] and [32]). Thus, the current study aims to fill the gap in the field by proposing a convergent parallel mixed-methods approach that begins, as an internal step, with a qualitative thematic analysis method and an NLP-based quantitative method (used independently), followed by training an unsupervised clustering model for integrating the internal analysis outputs for deeper insights in order to provide a deep understanding of the phenomena of viral YouTube videos in Saudi Arabia.

III. Research Methodology

The original work carried out in this paper was to better understand the rapid spread of viral YouTube videos in Saudi Arabia. We have considered a variety of video’s categories and qualitative themes as input factors for our experiment that conducted on a dataset collected from the top 13 viral videos, trended between 2016 and 2017 as reported in *Think-with-Google*[1]

Through the stages of this study, we have investigated the importance of sentiment analysis and mining opinions from YouTube comments, which allows us to classify the viewer’s

[1] Think-with-**Google**-https://www.thinkwithgoogle.com/intl/en-145/perspectives/local-articles/youtube-and-search-online-trends-mena-2016/

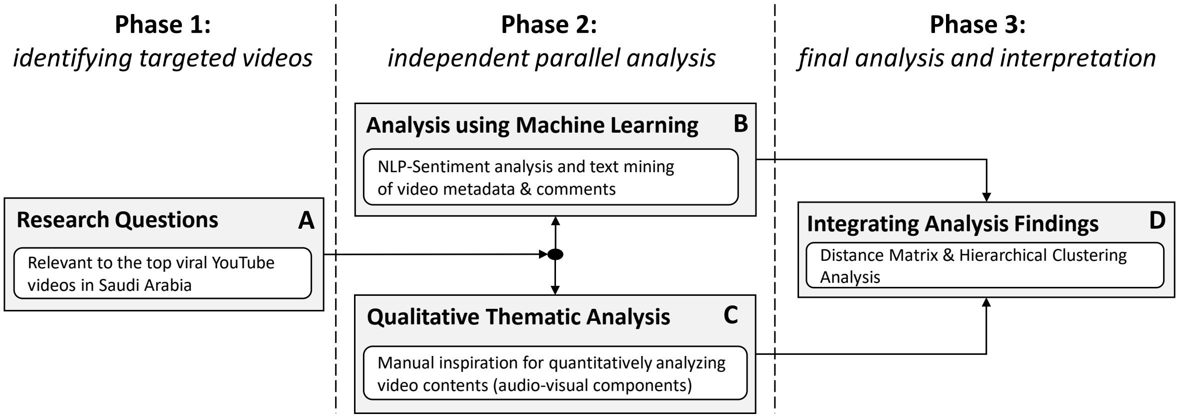

Phase 1:
identifying targeted videos
Phase 2:
independent parallel analysis
Phase 3:
final analysis and interpretation
Analysis using Machine Learning
B
NLP-Sentiment analysis and text mining of video metadata & comments
Research Questions
A
Relevant to the top viral YouTube videos in Saudi Arabia
Integrating Analysis Findings
D
Distance Matrix & Hierarchical Clustering Analysis
Qualitative Thematic Analysis
C
Manual inspiration for quantitatively analyzing video contents (audio-visual components)

Fig. 1. Our proposed convergent parallel mixed-methods approach

concerns and behaviors in conjunction with video's themes. By considering timestamps as an additional dimension, we attempt to anticipate the future shift of community concerns in social circles.

The convergent parallel mixed-methods approach, presented in this paper, integrates a qualitative thematic analyses method for analysing video content view with a quantitative method of NLP-mining opinion for investigating viewer's textual comments. The outcomes from these independent methods (i.e., qualitative and quantitative methods) are then integrated and fed into our an unsupervised machine learning (i.e., Hierarchical Cluster Analysis) model for comprehensive understanding and more accurate predictions. The overall flow proposed approach is described in a three phases, illustrated in Fig. 1. In the rest of this section, we first discuss our data collection methodology, including the selection criteria of YouTube videos for experiments as shown under *Phase 1*. We next introduce our analysis methods for both viewer's comments and video's contents for answering our research questions, see *Phase 2* and *Phase 3* of Fig. 1.

A. Acquiring Data for Experiments

YouTube is the second-most popular video-sharing website in the world, according to Alexa website[2]. It provides an official API[3] Services to access and fetch specific data that are available under their authorisation credentials. The publicly available data (i.e., free to fetch with restrictions) include general video meta-data, comments thread, limited user profile details, etc. We have developed a Java application with a mySQL database as a back-end to fetch/store only available public data.

We crawled all obtainable data related to those top 13 trended YouTube's videos in Saudi Arabia[1], uploaded/posted within a one-year timeframe (i.e. between 2016 and 2017). The gathered datasets includes more than $51,697$ comments and all the available details about reviewer profiles, such as location and used devises for posting comments. Critical demographic variables such as user-age and gender are unfortunately not available for public use, and hence, we had to implement our own classifier to predict user-genders from user-nicknames. Statistical summary of the collected datasets for our experiments in this research is given in Table I.

TABLE I. STATISTICAL SUMMARY OF THE DATASETS GATHERED FROM THE TOP 13 TRENDED YOUTUBE'S VIDEOS IN SAUDI ARABIA BETWEEN 2016 AND 2017.

#Vid.	Video ID	Viewers	Comments	like	dislike
1	1rUn2j1hLOo	11134481	14672	107259	27645
2	U62F_sl-D	2064982	3315	14811	3966
3	tE22WlRdEek	3979358	1810	6847	3106
4	wHggs-hE16M	1722002	3774	19500	1882
5	3QS7j-jDATE	335536	216	411	105
6	lxp-HDSARXs	2515276	5584	39590	1701
7	1yVWXXWwgnM	3353240	7780	43852	6956
8	5U02EzUWDmc	3281900	2406	68162	8534
9	oIHuAwYLW-U	10770907	3231	223952	17345
10	gOOOhdXT6QU	1598790	3471	49100	4061
11	Bzveyqagqeo	565006	876	2104	1015
12	HLX6D1jDzCg	499442	852	3338	1173
13	NHkCN058yFE	1095252	3710	8239	2621
Total		**42916172**	**51697**	**587165**	**80110**

B. NLP-Sentiment Analysis for Classifying Textual Comments

As the standard YouTube's API does not provide sentiment information correlated with each posted comment, we implement our own multi-classes sentiment classification algorithm for text written in Arabic. Our sentiment classifier algorithm is modelled using an optimised version of *bag-of-words* approach and analyses deeply sentiment scores in five-pole scale (i.e. *Positive, Negative, Mixed, Criticism, Neutral*) taking into consideration their polarities. The bag-of-words approach is popular in natural language processing, which is a machine learning method of feature extraction with textual data [35]. Rather than measuring only the presence and/or frequency of known words for a given textual comment, we also consider the sentiment score of each matched word from our predefined lexicon dictionary. We build a rich Arabic lexicon dictionary that includes more than $72,000$ sentimentally classified units, some of them have been extracted from publicly available datasets such as SemEval [36, 37] and from review repositories of some domains[4], including Movies, Hotels, Restaurants and

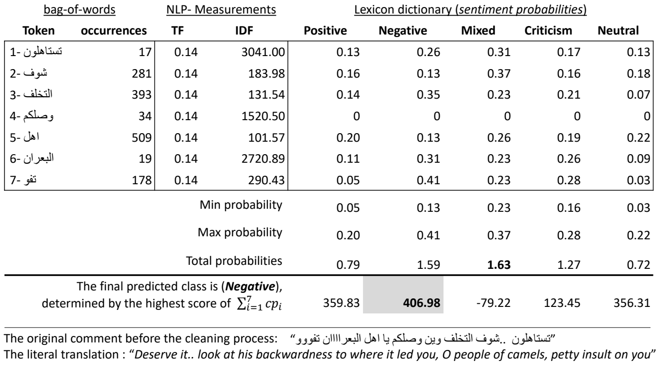
[2] Alexa Internet, Inc June 2019. https://www.alexa.com/siteinfo/youtube.com

[3] YouTube Application Program Interface (API) Services - https://developers.google.com/youtube/

[4] Large Multi-Domain Resources for Arabic Sentiment Analysis - https://github.com/hadyelsahar/large-arabic-sentiment-analysis-resouces

Products [38]. In principle, these units have been generated by mining varieties of Arabic texts that are currently in use, and the average of their accuracy is approximately 72%.

Moreover, our text classifier algorithm allows performing a detailed analyses of viewer's comments by predicting user-genders as well as classifying comments into another three high level categories (i.e. *Information, Conversation, Non-response comments*) using a specific predefined keywords. In this paper, these high level categories, introduced and explained in [39], could give influential facts that help in understanding the currently dominated phenomena of Saudi society.



bag-of-words		NLP- Measurements		Lexicon dictionary (*sentiment probabilities*)				
Token	occurrences	TF	IDF	Positive	Negative	Mixed	Criticism	Neutral
1- تستاهلون	17	0.14	3041.00	0.13	0.26	0.31	0.17	0.13
2- شوف	281	0.14	183.98	0.16	0.13	0.37	0.16	0.18
3- التخلف	393	0.14	131.54	0.14	0.35	0.23	0.21	0.07
4- وصلكم	34	0.14	1520.50	0	0	0	0	0
5- اهل	509	0.14	101.57	0.20	0.13	0.26	0.19	0.22
6- البعران	19	0.14	2720.89	0.11	0.31	0.23	0.26	0.09
7- تفو	178	0.14	290.43	0.05	0.41	0.23	0.28	0.03
			Min probability	0.05	0.13	0.23	0.16	0.03
			Max probability	0.20	0.41	0.37	0.28	0.22
			Total probabilities	0.79	1.59	**1.63**	1.27	0.72
			The final predicted class is (***Negative***), determined by the highest score of $\sum_{i=1}^{7} cp_i$	359.83	**406.98**	-79.22	123.45	356.31

The original comment before the cleaning process: “تستاهلون ..شوف التخلف وين وصلكم يا اهل البعراااان تفووو”
The literal translation : *“Deserve it.. look at his backwardness to where it led you, O people of camels, petty insult on you”*

Fig. 2. A self-explanatory example for analysing a textual Arabic comment, represented by a two-dimensional array-like structure: *bag-of-word* across *lexicon dictionary*. The latter includes five sentiment probabilities for each word. Here, we should notice that the negative values (i.e. see -79.22) when calculating the total score $\sum_{i=1}^{bag_{size}} cp_i$, using Equation 3, can be a result of not finding tokens in the dictionary, such as the token number 4.

```
Algorithm 1 Creating a bag of Arabic words Algorithm.
Inputs: $Comments = \{c_1, \cdots, c_k\}$, $k \in \mathbb{N}$: a set of all extracted comments from the datasets.
Outputs: $Bag$: a set consisting of a cleaned bag of Arabic words, such that each word $t$ has a numerical attribute $t_{count}$ for holding the number of comments the $t$ appear in.
Begin
  $Bag := \emptyset$ : initialising the empty bag set for creating distinct words.
  for each $c_i$ posted comment $\in$ Comments do
    $t_{cleaned} \leftarrow$ clean ($c_i$) : remove all non-Arabic characters, conjunctions, punctuation, and repeated stressing characters from $c_i$ except empty spaces.
    $t_{tokenized} \leftarrow$ Tokenize ($t_{clean}$) : tokenizing the passed cleaned text by splitting it on single spaces.
    for each $t_i$ a cleaned token $\in t_{tokenized}$ do
      if $t_i \notin Bag$ then
        $Bag \leftarrow t_i$ : append the token $t_i$ to the list $Bag$.
      if exist_and_first_count ($c_i$, $t_i$) then
        count how many comments the $t_i$ appear in Comments, i.e., at most once for each $c_i$.
        set $t_{i_{count}} = t_{i_{count}} + 1$
  return $Bag$.
End
```

```
Algorithm 2 Lexicon-based Bag of Words Sentiment Classifier.
Inputs: $Lex$ is a lexicon dictionary
        $Bag$ is the created bag of word from the Algorithm 1
        $tc$ and $pc$ are the total number of comments and a specific posted-comment respectively.
Outputs: $S_{class}$ is the classified class that has the maximum sentiment probabilistic scores from the five-pole scale (i.e., Positive, Negative, Mixed, Criticism, Neutral).
Begin
 1: $dataFrame$ = makeMatrix ($Bag$, $Lex$)
 2: $t_{cleaned} \leftarrow$ clean ($pc$)
 3: $t_{tokenized} \leftarrow$ Tokenize ($t_{cleaned}$)
 4: for each $t_i$ a cleaned token $\in t_{tokenized}$ do
 5:   if $t_i \in dataFrame[Bag]$ then
 6:     $TF \leftarrow dataFrame$.TF ($t_i$, $pc$) : compute Term Frequency using Equation 1
 7:     $IDF \leftarrow dataFrame$.IDF ($t_i$, $tc$) : compute Inverse Document Frequency using Equation 2
 8:     $dataFrame[t_i][Lex]$.sentimentScores ($TF * IDF$): compute the sentiment score for each row in $Lex$ according to their probabilities using formula Equation 3.
 9: $S_{class} \leftarrow dataFrame$.maxSentimentScore() : summing the total sentiment scores for each class in $Lex$ and then returns the class with the highest value.
10: return $S_{class}$
End
```

The proposed algorithms are given explicitly in (Algorithms 1 and 2). Given a broad set of textual comments, our approach begins by generating a bag of Arabic word using Algorithm 1 from all observed comments. We then generate a data-frame, representing a two-dimensional array-like structure, by mapping each token (word) from the bag with our predefined lexicon dictionary. Here, all columns are vectors of equal length, such that the first two vectors contain token-values and their occurrences respectively. The followed vectors correspond to measurements of the sentimental classes, obtained from our lexicon dictionary, see Figure 2 for illustration with a self-explanatory example. The generation of the data-frame is stated in Algorithm 2, see the first line.

We used Term-Frequency (TF) and Inverse-Document-Frequency (IDF) formulas for assessing how important a token is to a posted comment in corpus. These two statistical formulas, see Equation 1 and 2 are well-known measurements in text mining and information retrieval [40]. TF gives a scoring weight for each token in a document (i.e., how frequently a word appears in a comment), expressed as follows:

$$TF(t, c) = \frac{f_{t,c}}{c_{count}} \tag{1}$$

where $f_{t,c}$ is the number of times the token (or word) t appears in the posted comment c, and c_{count} is the total number of tokens. Whereas, IDF measures the score of how important a token is across documents (i.e., all observed comments), calculated by (2):

$$IDF(t, C) = \log \frac{C_{count}}{1 + |\{c \in C : t \in c\}|} \tag{2}$$

where C_{count} is the total number of extracted comments, and $|\{c \in C : t \in c\}|$ is the number of posted comments that the

TABLE II. COMMENTS CLASSIFICATION RESULTS AND THE PERCENTAGE OF IRRELEVANT COMMENTS (INCLUDING ADS) FOUND IN EACH VIRAL VIDEO

	Sentiment Classification					Keyword Classification			Ads and
#Vid.	Positive	Negative	Mixed	Criticism	Neutral	Info.	Conv.	Non-response	**irrelevant.**
1	1942	3056	1405	1125	7144	5023	2961	5817	40%
2	283	782	215	266	1769	1032	777	1390	42%
3	105	563	132	141	869	564	311	722	40%
4	861	953	745	342	873	1074	506	748	20%
5	25	69	19	23	80	74	71	67	31%
6	879	1576	939	555	1635	2179	975	1355	24%
7	603	2173	526	521	3957	2159	1104	3432	44%
8	341	709	336	413	607	881	106	482	20%
9	480	711	324	370	1346	1126	36	988	31%
10	508	703	330	298	1632	1265	561	1187	34%
11	80	194	74	155	373	238	228	309	35%
12	65	253	85	70	379	386	209	272	32%
13	304	1011	272	227	1896	878	764	1596	43%

TABLE III. RESULTS OF GENDER DETERMINATION IN EACH VIRAL VIDEO AND ESTIMATED RATE OF COMMUNICATION BETWEEN COMMENTERS

	Inferred Gender			
#Vid.	Male	Female	Unknown	**Interactions**
1	6956	3589	4127	20%
2	1607	813	895	23%
3	997	398	415	17%
4	2194	761	819	13%
5	121	43	52	33%
6	3294	1141	1149	17%
7	4182	1710	1888	14%
8	1236	544	626	4%
9	1645	745	841	1%
10	1920	787	764	16%
11	485	189	202	26%
12	464	193	195	25%
13	1850	862	998	21%

token t appears in it. We calculate *TF* and *IDF* for each token t in the posted comment c, see lines (4-7) of Algorithm 2. In line (8), we rescale data values in vectors that only correspond to the probabilities of the sentimental classes (i.e., columns with the header names: *Positive, Negative, Mixed, Criticism, Neutral* in Figure 2) using what so-called *feature scaling* multiplied by the calculated rates of token's importance $TF * IDF$ for each token t. To be more precise, rescaling these data values for the probabilities of each sentimental class is expressed as follows:

$$cp_t = TF_t * IDF_t \frac{cp_t - cp_{min}}{cp_{max} - cp_{min}} \tag{3}$$

where cp_{max} and cp_{min} are determined vertically in vectors that hold sentimental token's probabilities. Finally, the predicting sentimental class for c, stated in line (9) see Algorithm 2, is chosen according to the highest total scores of their $\sum_{i=1}^{b_{size}} cp_i$, where b_{size} is the size of the bag. Consider the example shown in Figure 2, the chosen sentimental class, the algorithm will classify the mentioned impolite comment to be *Mixed* in accordance with the total probabilities (i.e., 1.63). However, our optimised solution gives more precise classification as it takes into consideration the rates of token's importance, see the correct prediction by choosing *Negative* with total score of 406.98.

The same algorithms (i.e., Algorithms 1 and 2) are applied for classifying comments into three high level categories (i.e, *Information, Conversation, Non-response comments*), but with using different lexicon dictionary that is manually defined. Here, we carefully collect a large set of keywords that are often used in each category. For instance, comments that consist of WH-questions (as predefined keywords) at the beginning will likely be classified into *Information* category [39]. Furthermore, we have a rich database dictionary of male and female names, and we use it for predicting user-genders from user-nicknames.

Tables II and III show the generated predictions when applying our Algorithms 1 and 2 on the collected data, summarised in Table I.

C. Manual Inspiration for Quantitatively Analysing YouTube Content View

To our knowledge, there has been no idealistic method for performing video content analysis directly at the visual level. Accordingly, we have implemented a generic subjective method of interpretations by a panel of three reviewers, including the authors of this paper, moderately related to QualCA research method [41]. In essence, this subjective method involves three core independent phases:

1) The identification of the (global) most expressive themes and video categories that characterise the intentions deduced from the audio and/or visual components of video contents.
2) The coding frame, formulated in a two-dimensional thematic vector that maps the identified themes th_i with each observed video vid_i by a five-level Likert scale (i.e., from 1 to 5) [42].
3) Checking the validity of the constructed thematic vectors.

The selected 13 videos were distributed to the authors of this paper individually, and they were instructed to identify the global themes depending on what they observed in the video, regarded as a whole. Subsequently, the authors have held several remote meetings to unify all the agreed themes embedded in the video contents, wrapped into 10 distinct themes, described in Table IV. This phase was carried out during the month of January 2018. In the coding phase, the authors were requested individually to re-observe each video and scale all the 10-themes. To tackle the conflicting problems in scaling the same theme th_i vs. vid_i by the authors, the average scales has been calculated, and then rounded to the nearest integer. After that the authors have sent the constructed thematic vectors (i.e., represented as a table shown in Table IV) to a panel of three reviewers for checking and validating the identified list of themes as well as the coding scales for each video, and no critical comments were noticed.

TABLE IV. RESULTS OF MEASURING THE 10-THEMES FOR EACH CONTENT, INCLUDING AUDIO-VISUAL COMPONENTS, OF THE SELECTED 13 YOUTUBE VIDEOS, BASED ON A SCALE FROM 1 TO 5

	#Video ID												
Themes	1	2	3	4	5	6	7	8	9	10	11	12	13
1	3	5	5	3	0	0	0	0	2	0	0	0	5
2	5	4	4	5	2	4	2	3	4	1	2	1	4
3	4	3	3	3	0	2	0	1	0	0	0	0	4
4	0	1	1	1	3	0	0	5	5	0	1	0	1
5	0	0	0	0	1	5	0	0	0	0	2	0	0
6	0	0	0	0	0	5	0	0	0	0	3	0	0
7	0	0	0	0	0	5	0	0	0	0	1	0	0
8	5	1	1	3	5	1	3	3	4	5	5	5	1
9	2	2	2	4	0	2	5	0	2	1	1	0	2
10	3	3	3	0	2	0	2	2	2	3	2	5	3
	1 Opposite sex **2** Social and political issues **3** Religious issues **4** Celebrities and figures scandal **5** Defending the country **6** Supporting leaders **7** Feeling proud of the country **8** Sarcastic **9** Traditions **10** Sport and Entertainment												

D. Unsupervised Learning Model for Integrating the Quantitative and Qualitative Findings

The third phase of our concept-level mixed-methods design, shown in section III, makes the quantitative and qualitative findings interdependent rationally. It involves the integration of the NLP-Sentiment analysis (cf. subsection III-B) and the thematic analysis (cf. subsection III-C) outputs as the centerpiece inputs to our unsupervised machine learning model. This unsupervised learning model gives a more in-depth insight into the relations between the quantitative and qualitative variables, allowing to better understand the nature of viral videos. The proposed model, in the third phase, is designed based on Distance Matrix[5] and Hierarchical Clustering [43]. In data mining, distance matrix is typically essential for building a hierarchy of clusters. Here, we consider *Cosine* similarity formula (i.e., usually used to measure the degree of angle between two variables) to generate our distance matrix. The formula is expressed by a dot product [44] as follows:

$$\text{Distance (A,B)} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}} \tag{4}$$

where A_i and B_i are pairwise vectors containing values from two compared variables (e.g., a sentiment type as a variable vs. a specific user-gender).

```
# importing the required libraries ..
import pandas as pd
from scipy.spatial import distance_matrix
from sklearn.cluster import
    AgglomerativeClustering
...
# Loading the dataset as a CSV format and
# computing the distance matrix using the '
    scipy' library.
dataSet = pd.read_csv('cleaned-dataset.csv')
distanceMatrix = pd.DataFrame(distance_matrix(
    dataSet.values, dataSet.values), index=
    dataSet.index, columns=dataSet.index)
...
# Generating a hierarchical clustering model
    using the 'sklearn' library.
# We use the distance matrix as input to train
    the model.
hierarchicalClustering =
    AgglomerativeClustering (affinity='Cosine',
    linkage='ward')
hierarchicalClustering.fit_predict(
    distanceMatrix)
...
```

Listing 1. Python code fragments for computing distance matrix and generating a hierarchical clustering model

To clarify more, we have implemented a Python script to integrate all the inferred information acquired during the second phase (i.e., presented in Table II, Table III and Table IV). In Listing 1, we give a descriptive code fragment for creating a distance matrix between the qualitative thematic variables across the quantitative variables. Additionally, we have used an existing interactive data analysis tool called Orange[6] (i.e., a visual Python programing language for data analysis) to generate a distance matrix and hierarchical clustering figures, see them in Figure 3 and Figure 4. In principle, both figures illustrate the relations between the qualitative thematic variables across the quantitative variables. However, Figure 4 divides relations at different levels represented as a tree structure. We expand this and discuss our findings in the discussions section.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

This paper is set out to investigate the types of comments posted on viral YouTube videos in Saudi Arabia, proposing a thematic classification schema to understand the rates of concern of social community in Saudi Arabia. Without paying to much attention to our technical contribution to this study, which includes implementations of our own optimised learning algorithms and metrics, the focus in this section is to give revealing insights into the figures reported in Figure 3 and Figure 4 from three perspectives that answer our research questions:

1) Exploring the categorisation and current concerns of commenters under our qualitative themes. Here, we deeply dig into what more or less concerned the commenters depending on their genders as well as the level of their interactions.
2) Understanding what thematic categorisations are more relevant in boosting the spread of videos.

[5] Distance Matrix is a mathematical square matrix that contains the numerical distances between the items in two-dimensional-array

[6] Orange tool is an interactive data analysis workflows https://orange.biolab.si/

Distance Matrix	Positive	Negative	Mixed	Criticism	Neutral	Information	Conversation	Non-response comments	Male	Female	Unknown Gender	User interactions	Ads and irrelevant comments
Positive									0.022	0.088	0.091	0.198	0.429
Negative									0.038	0.022	0.014	0.299	0.390
Mixed									0.041	0.096	0.104	0.190	0.404
Criticism									0.069	0.107	0.091	0.195	0.437
Neutral									0.069	0.014	0.022	0.363	0.247
Information									0.033	0.063	0.071	0.223	0.497
Conversation									0.110	0.063	0.099	0.481	0.242
Non-response comments									0.066	0.016	0.019	0.365	0.242
Male	0.022	0.038	0.041	0.069	0.069	0.033	0.110	0.066		0.036	0.044	0.255	0.437
Female	0.088	0.022	0.096	0.107	0.014	0.063	0.063	0.016	0.036		0.014	0.343	0.324
Unknown Gender	0.091	0.014	0.104	0.091	0.022	0.071	0.099	0.019	0.044	0.014		0.324	0.335
User interactions	0.198	0.299	0.190	0.195	0.363	0.223	0.481	0.365	0.255	0.343	0.324		0.338
Ads and irrelevant comments	0.429	0.390	0.404	0.437	0.247	0.497	0.242	0.242	0.437	0.324	0.335	0.338	
1 Opposite sex	0.467	0.343	0.479	0.471	0.297	0.480	0.377	0.297	0.348	0.344	0.340	0.474	0.313
2 Social and political issues	0.236	0.198	0.239	0.278	0.275	0.300	0.335	0.268	0.333	0.334	0.223	0.356	0.453
3 Religious issues	0.322	0.222	0.302	0.375	0.242	0.378	0.241	0.242	0.358	0.331	0.251	0.490	0.401
4 Celebrities and figures scandal	0.300	0.303	0.313	0.367	0.262	0.256	0.135	0.277	0.210	0.358	0.303	0.408	0.334
5 Defending the country	0.420	0.368	0.396	0.444	0.312	0.396	0.461	0.337	0.365	0.245	0.392	0.257	0.372
6 Supporting leaders	0.452	0.491	0.478	0.426	0.430	0.478	0.430	0.456	0.366	0.246	0.483	0.369	0.404
7 Feeling proud of the country	0.452	0.491	0.478	0.426	0.430	0.478	0.430	0.456	0.361	0.243	0.483	0.369	0.404
8 Sarcastic	0.415	0.291	0.395	0.416	0.314	0.425	0.341	0.314	0.327	0.343	0.301	0.373	0.402
9 Traditions	0.188	0.124	0.222	0.251	0.152	0.206	0.196	0.145	0.316	0.348	0.130	0.290	0.343
10 Sport and Entertainment	0.321	0.418	0.329	0.308	0.429	0.404	0.450	0.452	0.334	0.353	0.445	0.322	0.217

Fig. 3. Distance matrix, based on cosine similarity formula, representing the relations between qualitative thematic variables across the quantitative variables.

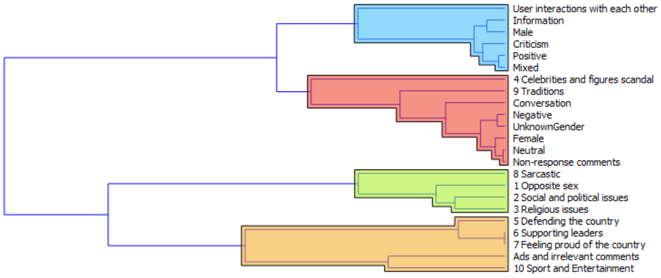


Fig. 4. Hierarchical clustering dendrogram, representing the relations between qualitative thematic variables across the quantitative variables.

3) Predicting the next shift wave of concerns of social commenters through observing all the event' times (i.e., time of posting or replying to a comment) on the timestamp.

After exploring the above-aforementioned perspectives in the next subsection, we discuss the threats that can impact the validity of our findings, and then we give a brief consideration regarding the ethical issues.

A. Understanding the Categorisations and Concerns of Saudi Society

Figure 5 shows four different distributions of our thematic categorisations, resulted by clustering our internal outputs (i.e., obtained after performing the qualitative and quantitative analysis parts independently). By taking a closer look at the mentioned figures as well as the disparity in percentages, one can observe, at a glance, the following points:

- The highest three categories in terms of social community concerns lie in (*Sport and Entertainment, Traditions* and *Sarcastic*), which constitute roughly half of the society's concerns in a total percentage of 49% (i.e. $18\% + 16\% + 15\%$), see (A) at the top left of Figure 5.
- The differences between males against females, as shown in (B), look slight by an average of about 11% except *Celebrities and figures scandal*, where they look more common among females than males by an approximate of 26%. This result is in line with the clustering dendrogram, presented in Figure 4. Here, the clustering figure gives different analytical readings, one of which is the overall behaviors of males against females. Roughly speaking, the produced clusters indicate that males appear more involved in making *positive, mixed* and *criticism* comments than females. These comments seem associated with all categorical themes apart from *Traditions* and *celebrities and figures scandal*. In contrast, however, females tend to post more *negative* and *neutral* comments associated with only *traditions* and *celebrities and figures scandal* categories.
- Interaction between commentators and their responses to each other is high in issues related to (*Political*, or *Sarcastic issues*), and gradually decrease in the other categories. Unsurprisingly, this is visibly analogous with the high presence of irrelevant comments, which can be a result of the exploitation of advertising owners in these categories, see (C) and (D).
- No much attempt is made by the commenters to delve into and engage in issues related to (*Political, Religious*, or *Opposite Sex issues*). However, there appears an advertising focus on these categories, which could be the reason behind boosting the level of communications between commenters.

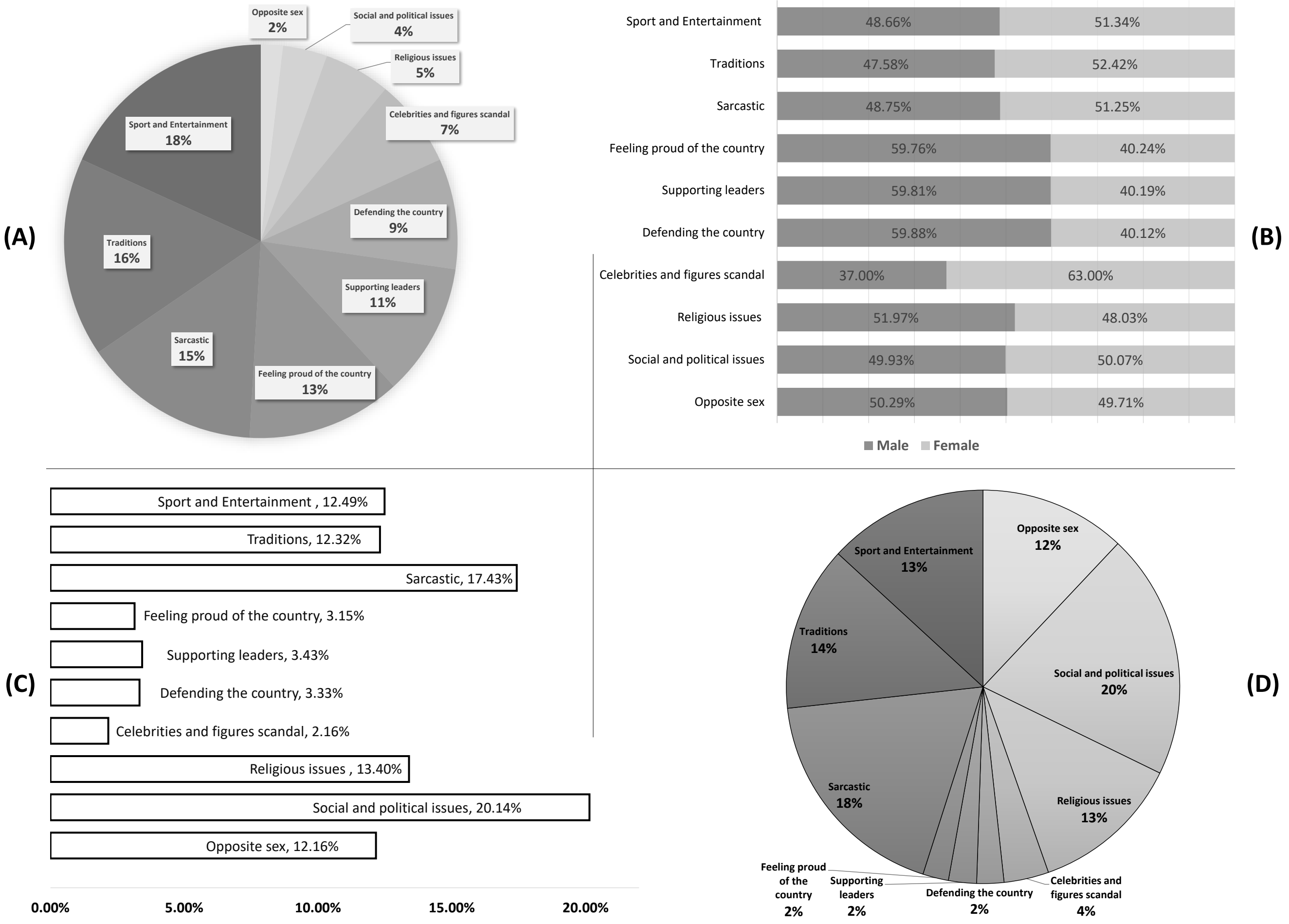


Fig. 5. Different distributions of our thematic qualitative categorisations. In (A), we show the distribution with percentages based on shared and/or posted comments. (B) shows the distribution in accordance with the percentage of male vs. female commentators. (C) focuses on the percentage of commentator-interactions with each other. The distribution, in (D), is determined based on the number of shared irrelevant comments, including textual Ads.

In order to entirely understand the phenomena of viral YouTube videos, one should collect data from several reliable resources that provide, e.g., tracing data of sharing video-links through external social networking platforms (or through existing mainstream media) or providing data that describe how much robot software tools being used for spreading video-links globally. Since such data are outside the scope of YouTube platform, let us assume *a hypothesis with a typical scenario where the genuine reason that led a particular video to go viral is just the content*. This trivial hypothesis simplifies our understanding of this phenomena by preciously examining one aspect (i.e., video's content in addition to its comments) while neglecting all other aspects that are difficult to obtain. In this context, we see the leading cause, confined to having an attractive positive or negative content, is the implication of what is in line with (1) the main interests of regular viewers or (2) with things that advertising organisations care about. Referring to such rational grounds, we figure out, from the results reported in Figure 5, that the prevalent categorical themes are *Traditions* and *Sarcastic*, thus supporting these categorises may contribute significantly to make an extremely viral video. Furthermore, what seems attracts the social community, in particular, is the promotion of entertainments and/or traditions issues associated with sarcastic or rude remarks. Advertisers, however, are keen to exploit contents correlated to politics, religion, opposite-sex issues and, in the meanwhile, surrounded by also rude remarks. Therefore, boosting these circumstances are likely the main reasons behind letting regular videos go viral.

Concerning our prediction for the changes in the distributed thematic categorisations, we have conducted a specific experiment to measure the changes. The concept of this experiment lies in adding event's times as an additional dimension to our dataset. To avoid the ambiguity, we firstly have broken the time-line down into several equal intervals, such that all our selected videos were accessible on-line during the first interval. Then, for each interval i, we generate a distance matrix by extracting comments, posted within i, and analysing them using our sentiment classifier. In (A) of Figure 6, we describe how the classified comments are fluctuated over time. By computing the generated set of distance metrics, using Forecast and Trendline equation[7], we were able to estimate the percentages of the change in each category, see (B) of Figure 6. This figure here reports that the changes, whether up or down,

[7] Forecast and Trendline are popular equation in MS-Excel tool.

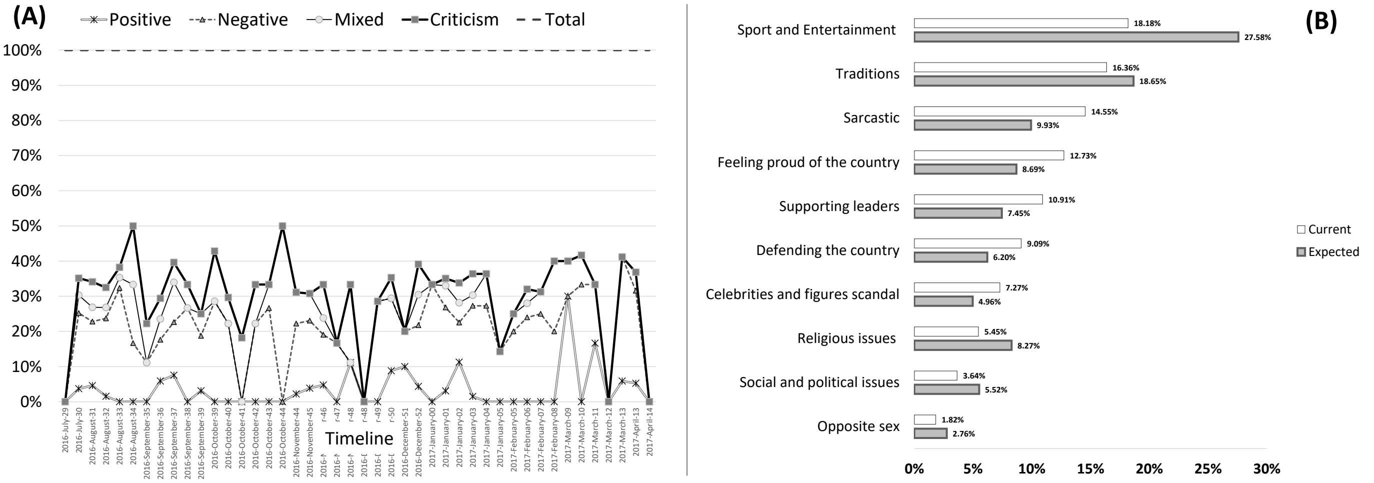


Fig. 6. Predicting the change in the current categorisations and concerns of social community. In (A), we describe the classified comments sentimentally for the first video (shown in Table I) over time. (B) illustrates the predicted change in each thematic category

would be inconsiderable of around 6%, except *Sport and Entertainment* that is expected to get progressed by almost 10%.

B. Discussion and Threats to Validity

The feasibility of using our complex AI-based approach in analysing the behaviors of YouTube communities depends primarily on the quality of the collected data, and this fact probably applies to most of in-use machine learning solutions. Consequently, our thought here is that viral YouTube videos can be considered as a fertile place for extracting high-quality dataset, resulting in producing accurate readings after correctly conducting required analysis. Concerning the soundness of our experiment, we discuss the threats that can impact the internal as well as the external validity of our results.

In internal validity (i.e., related to aspects that could have affected our finding), the threats may include (1) inaccurate predictions by our sentiment classifier, and (2) the improper use of cosine metric for generating distance matrices and hierarchical clustering (i.e., different metric may fit better in our approach such as *Manhattan* or *Euclidean*). For inaccurate predictions issue, we are not claiming that our sentiment classifier would give 100% correct predictions (no text-classifiers could reach this percentage), but accepting a specific prediction would often be based on a predefined threshold for a particular domain. The threshold considered in this paper is relatively close to the lowest accuracy of our lexicon units (i.e., about 63%) as we did exclude all lexicon units that have poor accuracy. This mean, the accuracy of our predictions should be above 63%, and such percentage is acceptable from the author's point of view. Regarding the use of cosine formula, intuitively using different formula will generate different results. However, cosine metric has been widely used for measuring preciously lexical similarity, and it is a typical metric for examining short text [45].

Threats to external validity investigate the scope of generalising the research findings. Here, a potential threat is represented by having incomplete data, collected from a limited number of (13) videos. While our approach deals with a single social networking platform (YouTube) in collecting data, there still relevant data left unconsidered, e.g., data from other social networking platforms as well as from chatbots software tools. However, as explained in the previous subsection, obtaining such data from external resources (i.e., outside the scope of YouTube platform) is not possible. Hence, we attempted to apply a robust and sophisticated research methodology using unsupervised machine learning for in-depth analysis and understanding. Besides, we are aware that our findings are based on a small number of carefully selected viral videos, but for ensuring a proper generalisation of our findings, we have extracted all shared comments (i.e., more than 51, 697 comments, see the details in Table I) within a one-year timeframe.

C. Ethical Considerations

Emotional feeling is individual privacy, and mining such individual privacy of a particular person evokes a significant concern regarding ethic legitimately. As intelligent machine learning systems are becoming more powerful and superior at understanding a human conversation, and their relationships, they could go beyond human ability in revealing their privacies, and hence raising critical questions to be addressed around security/privacy [46]. Technically speaking, mining what people express emotionally in the virtual social media worlds, as conducted in our experiments, is prone to random errors in disclosing the reality of the physical world. This means the predicted information, by mining algorithms, is not highly reliable and, therefore, could result in making ill-informed decisions.

Text mining and sentiment analysis approach on public resources of social media, as a knowledge-driven technique, are meant to give high societal level analytics. Despite this fact, our proposed approach is not designed to support oppressive regimes for identifying dissents and/or applying censorship. In this research, the collected datasets from YouTube's API are public and do not contain any details related to the identity of commenters. However, we have no attempt to use the inferred information, such as user-genders, to evaluate the private intellectual orientations of commenters.

To the best of our knowledge, no standard ethical guidelines exist to be implemented during the development of an artificial intelligence tool. However, there appears a promising attempt, which is not finalised yet, by a research group called Partnership on AI (PAI) to study the regulations and create such important guidelines [47].

V. Conclusion

This paper contributes a convergent analysing approach that can be applied, with negligible customisation, to any social video platform for in-depth analyses and comprehension. The principle underlying this approach depends on an unsupervised machine learning technique that integrates the internal outputs, obtained by applying qualitative and quantitative methods independently. For the latter method, we have introduced an optimised version of a well-known Bag of Words algorithm to sentimentally classify any given Arabic text into a five-pole scale using a rich lexicon dictionary. Our work also rationalised the importance of artificial intelligence (including NLP and machine learning) when dealing with a complex dataset that requires text mining analysis or analysing user behaviors.

We have empirically analysed $51,697$ comments, left on 13 trended YouTube videos along with their contents, for studying the phenomena of virality in Saudi Arabia. One of our main findings revealed that boosting entertainments, traditions, politics, and/or religion issues when making a video, that is associated in somehow with sarcastic or rude remarks, is likely the preeminent impulse for letting a regular video goes viral.

In the future, we intend to further optimise our parallel mixed-methods by semi-automating all the internal parts in a web-based application. We will be investigating on also optimising our sentiment classifier by taking into consideration the linguistic structure and grammar of texts.

References

[1] G. Feroz Khan and S. Vong, “Virality over youtube: an empirical analysis,” *Internet research*, vol. 24, no. 5, pp. 629–647, 2014.

[2] E. Botha, “A means to an end: Using political satire to go viral,” *Public Relations Review*, vol. 40, no. 2, pp. 363–374, 2014.

[3] K. Nahon and J. Hemsley, *Going viral*. Polity, 2013.

[4] R. A. STAN and C. Ana, “Emotions–drivers of online virality content characteristics of viral blog articles in romania,” *Local versus Global*, p. 694, 2015.

[5] R. Wang, W. Liu, and S. Gao, “Hashtags and information virality in networked social movement: Examining hashtag co-occurrence patterns,” *Online Information Review*, vol. 40, no. 7, pp. 850–866, 2016.

[6] M. Castells, “Networks of outrage and hope: Social movements in the internet age polity press,” 2012.

[7] A. J. Mills, “Virality in social media: the spin framework,” *Journal of public affairs*, vol. 12, no. 2, pp. 162–169, 2012.

[8] A. L. Montgomery, “Applying quantitative marketing techniques to the internet,” *Interfaces*, vol. 31, no. 2, pp. 90–108, 2001.

[9] S. Alhabash and A. R. McAlister, “Redefining virality in less broad strokes: Predicting viral behavioral intentions from motivations and uses of facebook and twitter,” *new media & society*, vol. 17, no. 8, pp. 1317–1339, 2015.

[10] M. Guerini, C. Strapparava, and G. Ozbal, “Exploring text virality in social networks,” in *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[11] S. Alhabash, J.-h. Baek, C. Cunningham, and A. Hagerstrom, “To comment or not to comment?: How virality, arousal level, and commenting behavior on youtube videos affect civic behavioral intentions,” *Computers in human behavior*, vol. 51, pp. 520–531, 2015.

[12] R. E. Guadagno, D. M. Rempala, S. Murphy, and B. M. Okdie, “What makes a video go viral? an analysis of emotional contagion and internet memes,” *Computers in Human Behavior*, vol. 29, no. 6, pp. 2312–2319, 2013.

[13] K. Nelson-Field, E. Riebe, and K. Newstead, “The emotions that drive viral video,” *Australasian Marketing Journal (AMJ)*, vol. 21, no. 4, pp. 205–211, 2013.

[14] Q. Bai, Q. V. Hu, L. Ge, and L. He, “Stories that big danmaku data can tell as a new media,” *IEEE Access*, vol. 7, pp. 53 509–53 519, 2019.

[15] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti, “Characterizing and modelling popularity of user-generated videos,” *Performance Evaluation*, vol. 68, no. 11, pp. 1037–1055, 2011.

[16] T. Broxton, Y. Interian, J. Vaver, and M. Wattenhofer, “Catching a viral video,” *Journal of Intelligent Information Systems*, vol. 40, no. 2, pp. 241–259, 2013.

[17] L. Jiang, Y. Miao, Y. Yang, Z. Lan, and A. G. Hauptmann, “Viral video style: A closer look at viral videos on youtube,” in *Proceedings of International Conference on Multimedia Retrieval*. ACM, 2014, p. 193.

[18] H. Pinto, J. M. Almeida, and M. A. Gonçalves, “Using early view patterns to predict the popularity of youtube videos,” in *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013, pp. 365–374.

[19] D. A. Shamma, J. Yew, L. Kennedy, and E. F. Churchill, “Viral actions: Predicting video view counts using synchronous sharing behaviors,” in *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[20] A. Susarla, J.-H. Oh, and Y. Tan, “Influentials, imitables, or susceptibles? virality and word-of-mouth conversations in online social networks,” *Journal of Management Information Systems*, vol. 33, no. 1, pp. 139–170, 2016.

[21] A. Vaish, R. Krishna, A. Saxena, M. Dharmaprakash, and U. Goel, “Quantifying virality of information in online social networks,” *International Journal of Virtual Communities and Social Networking (IJVCSN)*, vol. 4, no. 1, pp. 32–45, 2012.

[22] R. Zhou, S. Khemmarat, L. Gao, J. Wan, J. Zhang, Y. Yin, and J. Yu, “Boosting video popularity through keyword suggestion and recommendation systems,” *Neurocomputing*, vol. 205, pp. 529–541, 2016.

[23] K. English, K. D. Sweetser, and M. Ancu, “Youtube-ification of political talk: An examination of persuasion appeals in viral video,” *American Behavioral Scientist*, vol. 55, no. 6, pp. 733–748, 2011.

[24] K. Nahon and J. Hemsley, *Going viral*. Polity, 2013.

[25] R. Miller and N. Lammas, “Social media and its implications for viral marketing,” *Asia Pacific Public Relations Journal*, vol. 11, no. 1, pp. 1–9, 2010.

[26] I. Mohr, “Going viral: An analysis of youtube videos,” *Journal of Marketing Development and Competitiveness*, vol. 8, no. 3, p. 43, 2014.

[27] D. Southgate, N. Westoby, and G. Page, “Creative determinants of viral video viewing,” *International Journal of Advertising*, vol. 29, no. 3, pp. 349–368, 2010.

[28] J. Hautz, J. Füller, K. Hutter, and C. Thürridl, “Let users generate your video ads? the impact of video source and quality on consumers’ perceptions and intended behaviors,” *Journal of Interactive Marketing*, vol. 28, no. 1, pp. 1–15, 2014.

[29] J. Huang, S. Su, L. Zhou, and X. Liu, “Attitude toward the viral ad: Expanding traditional advertising models to interactive advertising,” *Journal of Interactive Marketing*, vol. 27, no. 1, pp. 36–46, 2013.

[30] O. F. Koch and A. Benlian, “Promotional tactics for online viral marketing campaigns: how scarcity and personalization affect seed stage referrals,” *Journal of Interactive Marketing*, vol. 32, pp. 37–52, 2015.

[31] J. M. Leonhardt, “Tweets, hashtags and virality: Marketing the affordable care act in social media,” *Journal of Direct, Data and Digital Marketing Practice*, vol. 16, no. 3, pp. 172–180, 2015.

[32] E. Shehu, T. H. Bijmolt, and M. Clement, "Effects of likeability dynamics on consumers' intention to share online video advertisements," *Journal of Interactive Marketing*, vol. 35, pp. 27–43, 2016.

[33] C. Tucker, "Virality, network effects and advertising," *The Networks, Electronic Commerce, and Telecommunications*, 2011.

[34] H. S. Kim, "Attracting views and going viral: How message features and news-sharing channels affect health news diffusion," *Journal of Communication*, vol. 65, no. 3, pp. 512–534, 2015.

[35] A. Kao and S. R. Poteet, *Natural language processing and text mining*. Springer Science & Business Media, 2007.

[36] S. Kiritchenko, S. M. Mohammad, and M. Salameh, "Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases," in *Proceedings of the International Workshop on Semantic Evaluation*, ser. SemEval 16, San Diego, California, June 2016.

[37] S. K. Mohammad Salameh, Saif M. Mohammad, "Arabic sentiment analysis and cross-lingual sentiment resources," http://saifmohammad.com/WebPages/ArabicSA.html, 2019, [Online; accessed 1-September-2019].

[38] H. ElSahar and S. R. El-Beltagy, "Building large arabic multi-domain resources for sentiment analysis," in *Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II*, 2015, pp. 23–34. [Online]. Available: https://doi.org/10.1007/978-3-319-18117-2_2

[39] A. Madden, I. Ruthven, and D. McMenemy, "A classification scheme for content analyses of youtube video comments," *Journal of Documentation*, vol. 69, no. 5, pp. 693–714, 2013. [Online]. Available: https://doi.org/10.1108/JD-06-2012-0078

[40] A. Rajaraman and J. D. Ullman, *Data Mining*. Cambridge University Press, 2011, p. 1–17.

[41] B. M. Wildemuth, *Applications of social research methods to questions in information and library science*. ABC-CLIO, 2016.

[42] W. L. Neuman, "Social science methods: Quantitative and qualitative approaches," 2011.

[43] T. Hastie, T. Robert, and J. Friedman, "The elements of statistical learning: Hierarchical clustering," 2009.

[44] P. Dangeti, *Statistics for machine learning*. Packt Publishing Ltd, 2017.

[45] G. Sidorov, A. F. Gelbukh, H. Gómez-Adorno, and D. Pinto, "Soft similarity and soft cosine measure: Similarity of features in vector space model," *Computación y Sistemas*, vol. 18, no. 3, 2014.

[46] N. Hutchins, Z. Kirkendoll, and L. Hook, "Social impacts of ethical artifical intelligence and autonomous system design," in *2017 IEEE International Systems Engineering Symposium (ISSE)*, Oct 2017, pp. 1–5.

[47] "Tenets - partnership on ai," https://www.partnershiponai.org/tenets/, 2019, [Online; accessed 1-September-2019].

Predicting Students' Performance of the Private Universities of Bangladesh using Machine Learning Approaches

Md. Sabab Zulfiker[1], Nasrin Kabir[2], Al Amin Biswas[3], Partha Chakraborty[4], and Md. Mahfujur Rahman[5]
Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh[1,2,3,5]
Department of Computer Science and Engineering, Comilla University, Cumilla, Bangladesh[4]

Abstract—**Every year thousands of students get admitted into different universities in Bangladesh. Among them, a large number of students complete their graduation with low scoring results which affect their careers. By predicting their grades before the final examination, they can take essential measures to ameliorate their grades. This article has proposed different machine learning approaches for predicting the grade of a student in a course, in the context of the private universities of Bangladesh. Using different features that affect the result of a student, seven different classifiers have been trained, namely: Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression, Decision Tree, AdaBoost, Multilayer Perceptron (MLP), and Extra Tree Classifier for classifying the students' final grades into four quality classes: Excellent, Good, Poor, and Fail. Afterwards, the outputs of the base classifiers have been aggregated using the weighted voting approach to attain better results. And here this study has achieved an accuracy of 81.73%, where the weighted voting classifier outperforms the base classifiers.**

Keywords—Prediction; machine learning; weighted voting approach; private universities of Bangladesh

I. Introduction

Nowadays various statistical and machine learning algorithms are applied in different fields, such as marketing, health and medical issues, weather forecasting, socioeconomic behavior analysis, etc. It has emerged to educational data also. From the perspective of Bangladesh, the number of the private universities is much more than the number of the public universities. Currently in Bangladesh, there are one hundred and five private universities [1]. As a result, the number of students in the private universities is much higher than the public universities. Since we cannot imagine the development of the higher education except the development of the quality of the private universities, it is necessary to focus on the students of the private universities.

The databases of the different universities store a large volume of data. This data include the data of the students, teachers, and employees of the universities. By analyzing this data, different patterns can be derived which will be helpful to make decisions. Using diverse machine learning and data mining techniques on these data, many kinds of knowledge can be discovered and this knowledge can be used to predict the enrolment status of the students in a course, to detect illegal activities in the online examination and, to identify unusual marks in the result sheet, etc. [2]. Different statistical analysis and machine learning algorithms can be applied on the data of the students of the universities for predicting the grades of different courses that they have taken in their undergraduate level. There is a massive growth in the number of students who are getting admitted in different public and private universities of Bangladesh. A vast portion of these students can not gather proper skills and knowledge in their four years' tenure of the university life. Not only practical knowledge, a huge number of students come out of universities with low scoring results. As they lack both theoretical knowledge and practical knowledge, it becomes very hard for them to place themselves in job markets. If the students can predict their grades or results before their final examinations, they can take necessary actions to improve their results. Then the teachers can also identify which students are at risk and so they can guide the weak students properly and help them to recover [3]. For predicting the performance of the students predictive modeling can be used. Several methods can be used for building a predictive model like: classification, regression, categorization. Among these methods, classification has the most popularity [4].

After the accomplishment of this research, it will help to find out the different approaches to predict the students' final grades as well as determine the best approach for performing the prediction.

The main objectives of this research are: to predict the final grades of the students in a course using different machine learning algorithms, to forecast whether a student is at risk of failure in the final examination or not, and to compare the results of different machine learning algorithms for identifying which algorithm gives the best performance.

The residue of the paper is structured as follows: Section II describes the related works, Section III exhibits the entire methodology. The results are discussed in Section IV, the conclusion is discussed in Section V. Finally, Section VI represents the future works.

II. Related Works

Yadav and Pal [2] performed a study on predicting the results of the students of 1st year of Engineering. They collected data from the enrolment form which were filled by the students during their admission in VBS Purvanchal University, Jaunpur. With this dataset, they built models using different variations of the Decision Tree algorithm for classifying the students' performances in the year final examination of the first year. They showed that C4.5 obtained the highest accuracy of 67.78%.

Kabra and Bichkar [3] collected data from the entry form, filled by the students in an engineering college during the time of admission. Using J48 algorithm, they predicted the final grades of the first year students'. When they classified the results into three categories they gained an accuracy of 60.46% and in the case of classifying the results into two categories they gained an accuracy of 69.94%.

Kapur et al. [5] used various classification algorithms to classify the performance of the students into three categories: high, medium, and low. Their dataset included 480 entries with 16 attributes. Among these classifiers Random Forest showed the highest accuracy of 76.67%.

Liu and Zhang [6] gathered 210 records of the students. The dataset contained the marks of some major subjects and with this dataset, they trained C4.5 classifier for predicting whether a student would pass or fail.

Sweeney et al. [7] proposed a system for predicting the grades of the students for the next enrollment term. They applied two classes of methods: Simple Baselines and Matrix Factorization (MF) based methods. The lowest prediction error was achieved by the Factorization Machine (FM) Model of Matrix Factorization based methods.

Yadav et al. [8] performed a comparative study among the CART, C4.5, and ID3 algorithms for predicting the end semester marks. The dataset contained a variety of attributes like: marks achieved in the last semester, grades obtained in the class test, attendance marks, lab work performances, etc. They used WEKA explorer as the data mining tool.

Minaei-Bidgoli et al. [9] performed their study on LON-CAPA, which is an online education system. Firstly, for the classification purpose they used diverse base classifiers like: Parzen window, 1- Nearest Neighbor (1NN), K-Nearest Neighbor (KNN), Quadratic Bayesian Classifier, Multilayer Perceptron (MLP), and Decision Tree. For improving the accuracy, they also made use of a combination of the classifiers. Finally, for optimizing the accuracy of the combination of the classifiers they used Genetic Algorithm (GA). They found that Genetic Algorithm increased the accuracy by 10-12%.

Z. Iqbal et al. [10] found that the CGPA of a student in the degree program is high, if his university entry test score and HSSC (Higher Secondary School Certificate) score is high. They compared the performance of Restricted Boltzmann Machine (RBM), Matrix Factorization (MF), Collaborative Filtering (CF) and showed that RBM exhibited the best performance.

A comparative study between four distinct models: Stepwise Polynomial Regression, Linear Decision Rule, Linear Multiple Regression, and a simple Artificial Neural Network were proposed by Gorr et al. for predicting students' GPA [11].

Meier et al. [12] stated that the timely prediction of the final grade is also significant. So they proposed an algorithm that could not only predict the final grades but also performed timely prediction using previous performances of the students.

Jishan et al. [13] showed that preprocessing the data with the combination of Synthetic Minority Over-Sampling and Optimal Equal Width Binning significantly improves the accuracy of predicting students' final grades.

Socio-demographic data of over 450 students, which were collected during the time of enrollment at the Open Polytechnic of New Zealand were analyzed by Kovacic [14] for predicting students' success. For classification purposes he applied CHAID and CART algorithms and showed that CART transcended CHAID.

Hijazi and Naqvi [15] identified several factors that influenced the students' performance in the intermediate examination using simple linear regression. They found that class attendance, family income, mother's education, and study hours per day have a proportional relation with the student's performance, and mother's age has a reverse proportional relation with the result.

Mia et al. [16], proposed different machine learning techniques for predicting the registration status of the private university's students of Bangladesh. Among the different classifiers, Support Vector Machine outperformed all other classifiers and achieved an accuracy of 85.76%.

Biswas et al. [17] used diverse machine learning classifiers to predict the enrollment and dropout status in the post-graduation level. For this work, they collected the dataset from a renowned public university of Bangladesh. They computed the performance evaluation metrics for each of the classifiers. Finally, they found that the locally weighted learning outstrips the other classifiers.

III. Methodology

This section is divided into three subsections: data description, algorithms description, implementation procedures. The subsections are briefly described below.

A. *Data Description*

The dataset used in this study has been obtained from a reputed private university of Bangladesh. It contains the records of 400 students of diverse courses of different departments from Summer 2018 to Fall 2019. This research has been performed using eight attributes, among them only one attribute is the response variable and the other seven attributes are predictor variables. These variables are described below in details.

- ATTDM: This attribute depicts the attendance marks of a student.
- RTK: It represents whether a student has retaken the subject or not.
- APAQ: During the tenure of a single semester, a student has to give three quizzes in a particular course. This attribute portrays whether a student appeared in all the quizzes or not.
- AQM: The average of the obtained quiz marks is depicted by this attribute.
- MIDM: The obtained marks in the mid term examination is represented by this attribute.
- SUAS: This attribute confirms whether a student has submitted the assignment or not.

TABLE I. GRADING POLICY OF THE UNIVERSITY GRANTS COMMISSION (UGC) OF BANGLADESH

Marks out of 100	Letter Grade	Grade Point	Marks out of 100	Letter Grade	Grade Point
80-100	A+	4.00	55-59	B-	2.75
75-79	A	3.75	50-54	C+	2.50
70-74	A-	3.50	45-49	C	2.25
65-69	B+	3.25	40-44	D	2.00
60-64	B	3.00	0-39	F	0.00

TABLE II. VARIABLES FOR PERFORMING PREDICTION

Variable Name	Variable Type	Data Type	Possible Values
ATTDM	Predictor Variable	Real Number	0-7
RTK	Predictor Variable	Categorical	Y (Yes) N (No)
APAQ	Predictor Variable	Categorical	Y (Yes) N (No)
AQM	Predictor Variable	Real Number	0-15
MIDM	Predictor Variable	Real Number	0-25
SUAS	Predictor Variable	Categorical	Y (Yes) N (No)
PPRE	Predictor Variable	Categorical	Y (Yes) N (No)
FNLG	Response Variable	Real Number	0 (Fail) 1 (Poor) 2 (Good) 3 (Excellent)

- PPRE: Represents whether a student has performed the presentation or not.
  The above seven attributes are the predictor variables.
- FNLG: This is the only response variable. It depicts the final grade of a student after the final examination. The university follows the grading policy of University Grants Commission (UGC) of Bangladesh which is shown in Table I [18].

The final grades are categorized into four categories. If a student achieves A+, A or A- , then his grade is categorized into the category ‘Excellent’. B+, B and B- are considered as ‘Good’. The letter grades C+, C, D are categorized as ‘Poor’, and the grade F is considered as ‘Fail’. The possible values, data types, and variable types of different variables used in this research are shown in Table II.

B. Algorithms Description

The algorithms used in this research are described below in details.

1) Support Vector Machine (SVM): Support Vector Machine (SVM) tries to separate two classes using an optimal hyperplane [19]. It uses supervised learning. SVM works better, if the size of the data is small [20]. It attempts to make the decision boundary to such a degree that the partition between two classes is as broad as could reasonably be expected. To separate two classes, let’s assume we are given a training data set, $D = (x_1, C_1), (x_2, C_2), ..., (x_N, C_N)$ where x_i denotes input vector and C_i refers to the class label of the vector which could be specified as either positive or negative. For specifying any unspecified vector X, the condition is as follows:

$$f(X) = \sum_{i=1}^{N} a_i C_i (x_i^T X) + b \qquad (1)$$

Here, the nonzero coefficients are $a_i \ (i = 1, 2, ..., N)$ and the bias is represented by b [21].

2) Logistic Regression: The relationship between different variables are settled by regression analysis. If the relationship is linear, then Linear Regression analysis can be applied. But in the case of nonlinear relationship between the variables, we can’t apply Linear Regression and Logistic Regression can be introduced then. Logistic Regression is a generalized form of Linear Regression [22]. Consider the following equation for the Linear Regression:

$$y = \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + ... + \alpha_n Z_n \qquad (2)$$

Here, y is the response variable and $Z_1, Z_2, Z_3, ...Z_n$ are the predictor variables. By applying the sigmoid function on the equation, we can get the logistic function.

$$l = 1/[1 + e^{-(\alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + ... + \alpha_n Z_n)}] \qquad (3)$$

3) K-Nearest Neighbor (KNN): A simple, non-parametric supervised learning algorithm is K-Nearest Neighbor algorithm, which can be used for both regression and classification. Based on the feature similarity (e.g. distance function), all the available cases are stored and new cases are classified by it. The output is a class membership in KNN classification. A case is categorized by a predominance vote of its neighbors.

The case is allotted to the utmost common class among its K nearest neighbors. Various heuristic techniques can select the value of K (positive integer) in KNN method. The case will be assigned to the class of its nearest neighbor if K=1 [23]. Different distance functions like: Minkowski Distance, Manhattan Distance, Euclidean Distance are used in KNN algorithm. In this work, Minkowski Distance function has been used. The Minkowski Distance for two points U $(u_1, u_2, ..., u_n)$ and V $(v_1, v_2, ..., v_n)$ can be represented by the following equation, where q represents the order of the Minkowski Distance.

$$distance(U, V) = (\sum_{i=1}^{n} (|u_i - v_i|)^q)^{1/q} \tag{4}$$

4) Decision Tree: Decision Tree classification uses tree like structures. The internal nodes of the tree represent the conditions and the external nodes or the leaves represent the class labels. The branches from the internal nodes represent the outcomes of the tests or conditions. The decision of splitting the data is controlled by entropy and which can be defined by the equation below, where p_j is the probability of the j^{th} class.

$$E(S) = \sum_{j=1}^{c} -p_j \ \log_2 p_j \tag{5}$$

Different variations of Decision Tree are available as for instance: ID3, C4.5, CART, etc.

5) AdaBoost: AdaBoost stands for Adaptive Boosting classifier. A set of weak classifiers is combined into a strong one using this approach. Here, the following equation represents the classification using the AdaBoost algorithm.

$$F(t) = sign(\sum_{m=1}^{M} \theta_m f_m(t)) \tag{6}$$

Here, the m^{th} weak classifier is represented by f_m and θ_m represents the corresponding weight.

6) Multilayer Perceptron (MLP): Multilayer Perceptron is a form of feedforward neural network and it consists of multiple layers of neurons. A neuron of one layer interacts with the neurons of its adjacent layers through weighted connections though there exists no connection between the neurons of the same layer. Excluding the input and the output layers, the MLP has one or more hidden layers or intermediate layers [24]. The error of the k^{th} output node in the data point n can be represented by the equation below where d and c represent the actual and predicted values respectively.

$$e_k(n) = d_k(n) - c_k(n) \tag{7}$$

7) Extra Tree Classifier: A variant of Random Forest known as Extra Tree Classifier was first introduced by Geurts et al. [25]. Extra Tree Classifier differs from other tree based classifiers in such a way that it uses the entire learning sample for growing the trees and it chooses cut-points for splitting the nodes fully at random.

8) Weighted Voting Classifier: Voting Classifier is an approach for combining the outputs of different base classifiers as it is hard to identify a specific classification algorithm that gives the best accuracy on a certain data. Both homogeneous and heterogeneous models can be aggregated using the Voting Classifier. In the weighted voting approach, a weight or coefficient is assigned to each base classifier which is proportional to the base classifier's individual accuracy [26]. Consider $h_1, h_2, h_3, ...h_n$ are the outputs of n-different classifiers respectively and $s_1, s_2, s_3, ..., s_n$ are the assigned weights to each classifier, respectively, then the final output H of the Weighted Voting Classifier can be represented by the following equation.

$$H = s_1 * h_1 + s_2 * h_2 + s_3 * h_3 + ... + s_n * h_n \tag{8}$$

C. Implementation Procedures

The implementation procedures are illustrated in this section. To carry out the study, Python and Scikit-learn library have been used.

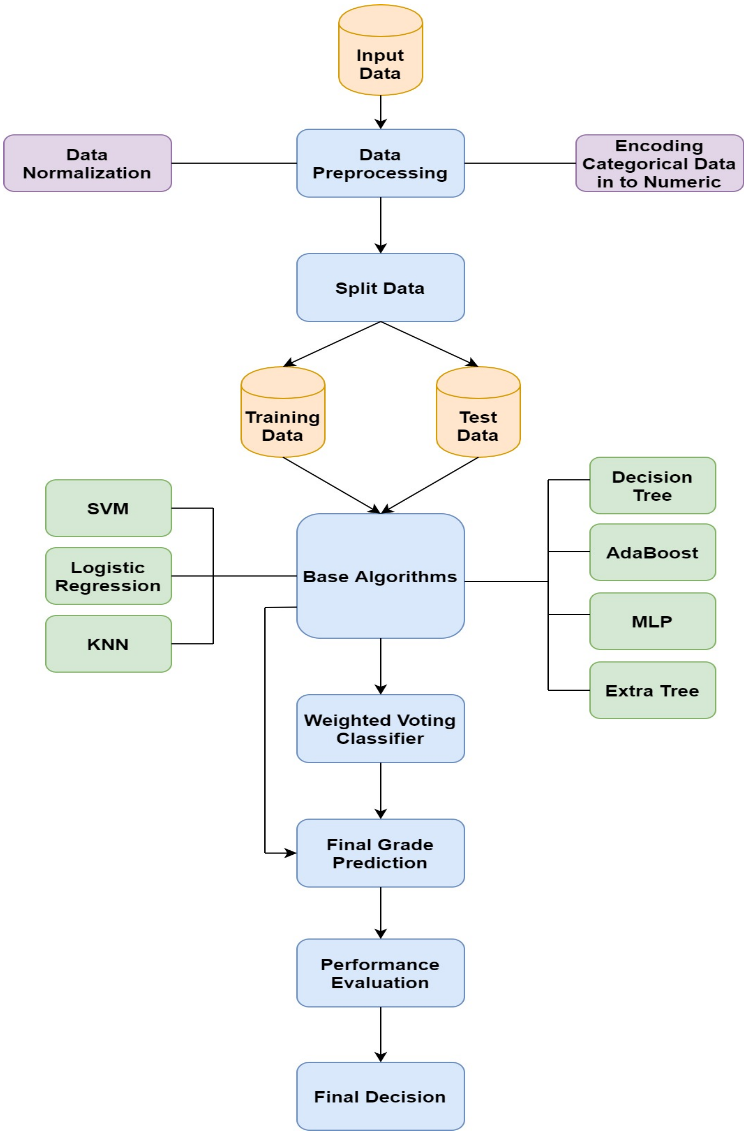


Fig. 1. Step by Step Procedures for Predicting Students' Final Grades

The graphical form of the stepwise procedures for predicting students' final grades is represented by Fig. 1. The details of Fig. 1 is depicted below.

1) Input Data: After collecting the data of 400 students via the Enterprise Resource Planning (ERP) system of the university, the task of inputting the data in the proposed system has been performed in this step.

2) Data Preprocessing: Data preprocessing step is categorized into two categories, namely: Data normalization and Encoding the categorical data into numeric data. In the collected dataset, the attendance marks range from 0 to 7, the average quiz marks range from 0 to 15 and the obtained mid term examination marks range from 0 to 25. Under the circumstances, these three predictor variables are in very different ranges. So, normalization of these three attributes has been performed. After the normalization procedure, the values of these three variables range from 0 to 1. There are some categorical data in the dataset. Algorithms like Decision Tree algorithm can work effectively with categorical data but most of the other algorithms give better performance while using numerical data instead of its categorical counterpart. Hence, the categorical data have been encoded into numerical data using the Label Encoding approach of the Scikit-learn library.

3) Data Splitting: Splitting the dataset follows the data preprocessing step. This step splits the dataset into training data and test data. In this work, 74% of data is used for training purposes and the rest 26% of data is used for testing.

4) Training and Testing using Base Classifiers: In this step, the seven base classifiers have been trained with the training data. And after training the classifiers, the prediction of the final grades of the students has been performed using the test data. Accuracy of each base classifier is also measured separately.

5) Aggregating the Outputs of the Base Classifiers: Eventually, using the Weighted Voting Classifier, this step aggregates the outputs of these seven base classifiers for achieving better performance.

6) Performance Evaluation: This step compares the performance of the base classifiers with the performance of the Weighted Voting Classifier. For evaluating the performance, five evaluation metrics: Accuracy, Precision, Recall, F-1 score and Area Under Curve (AUC) are determined.

7) Final Decision: According to the outcomes of the evaluation metrics, the best classifier for predicting the final grades of the students has been selected in this step.

IV. Results

For testing purposes, the records of 104 students have been used. Among these records, 38% records are actually classified

TABLE III. Confusion Matrices for Different Classifiers for Predicting Students' Grades

Classifiers Name	**Predicted→**	**Fail**	**Poor**	**Good**	**Excellent**
	Actual↓				
SVM	**Fail**	2	1	0	0
	Poor	3	29	5	0
	Good	0	4	10	10
	Excellent	0	1	1	38
Logistic Regression	**Fail**	2	1	0	0
	Poor	2	19	13	3
	Good	0	2	11	11
	Excellent	0	0	3	37
KNN	**Fail**	3	0	0	0
	Poor	3	30	4	0
	Good	0	5	16	3
	Excellent	0	1	11	28
Decision Tree	**Fail**	3	0	0	0
	Poor	4	25	8	0
	Good	0	2	19	3
	Excellent	0	1	7	32
AdaBoost	**Fail**	3	0	0	0
	Poor	3	19	14	1
	Good	0	1	15	8
	Excellent	0	0	4	36
MLP	**Fail**	2	1	0	0
	Poor	2	24	10	1
	Good	0	2	15	7
	Excellent	0	0	3	37
Extra Tree	**Fail**	3	0	0	0
	Poor	3	29	5	0
	Good	1	6	13	4
	Excellent	0	1	6	33
Weighted Voting Classifier	**Fail**	3	0	0	0
	Poor	3	29	5	0
	Good	0	2	19	3
	Excellent	0	0	6	34

TABLE IV. MEASURED RESULTS OF DIFFERENT CLASSIFIERS FOR PREDICTING STUDENTS' GRADES

Classifiers Name	Accuracy	Class Label	Precision	Recall	F-1 Score	AUC
SVM	75.96%	**Fail**	0.40	0.67	0.50	0.81
		Poor	0.83	0.78	0.81	
		Good	0.63	0.42	0.50	
		Excellent	0.79	0.95	0.86	
Logistic Regression	66.35%	**Fail**	0.50	0.67	0.57	0.76
		Poor	0.86	0.51	0.64	
		Good	0.41	0.46	0.43	
		Excellent	0.73	0.93	0.81	
KNN	74.04%	**Fail**	0.50	1.0	0.67	0.85
		Poor	0.83	0.81	0.82	
		Good	0.52	0.67	0.58	
		Excellent	0.90	0.70	0.79	
Decision Tree	75.96%	**Fail**	0.43	1.0	0.60	0.87
		Poor	0.89	0.68	0.77	
		Good	0.56	0.79	0.66	
		Excellent	0.91	0.80	0.85	
AdaBoost	70.19%	**Fail**	0.50	1.0	0.67	0.83
		Poor	0.95	0.51	0.67	
		Good	0.45	0.63	0.53	
		Excellent	0.80	0.90	0.85	
MLP	75%	**Fail**	0.50	0.67	0.57	0.81
		Poor	0.89	0.65	0.75	
		Good	0.54	0.63	0.58	
		Excellent	0.82	0.93	0.87	
Extra Tree	75%	**Fail**	0.43	1.0	0.60	0.85
		Poor	0.81	0.78	0.79	
		Good	0.54	0.54	0.54	
		Excellent	0.89	0.82	0.86	
Weighted Voting Classifier	81.73%	**Fail**	0.50	1.0	0.67	0.90
		Poor	0.94	0.78	0.85	
		Good	0.63	0.79	0.70	
		Excellent	0.92	0.85	0.88	

as “Excellent”, 23% records are actually classified as “Good”, 36% records are actually classified as “Poor” and the other 3% records are originally classified as “Fail”.

The confusion matrices of the result of this study using SVM, Logistic Regression, KNN, Decision Tree, AdaBoost, MLP, Extra Tree and Weighted Voting Classifier are presented in Table III.

The calculated Accuracy, Precision, Recall, F-1 Score and AUC (Area Under Curve) are shown in Table IV.

Table IV exhibits that Logistic Regression gives the lowest accuracy of 66.35%, where the Weighted Voting Classifier attains the highest accuracy of 81.73%. After the Weighted Voting Classifier, the second-highest accuracy of 75.96% is attained by the SVM and Decision Tree classifier jointly. The

TABLE V. MEASURED RESULTS OF DIFFERENT CLASSIFIERS USING TWO CLASS LABELS

Classifiers Name	Accuracy	Class Label	Precision	Recall	F-1 Score	AUC
SVM	92.31%	**Lower Order Grades**	1.0	0.80	0.89	0.90
		Higher Order Grades	0.89	1.0	0.94	
Logistic Regression	81.73%	**Lower Order Grades**	0.92	0.57	0.71	0.77
		Higher Order Grades	0.78	0.97	0.87	
KNN	90.38%	**Lower Order Grades**	0.86	0.90	0.88	0.90
		Higher Order Grades	0.94	0.91	0.92	
Decision Tree	83.65%	**Lower Order Grades**	0.81	0.75	0.78	0.82
		Higher Order Grades	0.85	0.89	0.87	
AdaBoost	83.65%	**Lower Order Grades**	0.85	0.70	0.77	0.81
		Higher Order Grades	0.83	0.92	0.87	
MLP	84.61%	**Lower Order Grades**	0.93	0.65	0.76	0.81
		Higher Order Grades	0.82	0.97	0.89	
Extra Tree	77.88%	**Lower Order Grades**	0.72	0.70	0.71	0.76
		Higher Order Grades	0.82	0.83	0.82	
Weighted Voting Classifier	93.26%	**Lower Order Grades**	1.00	0.82	0.90	0.91
		Higher Order Grades	0.90	1.00	0.95	

third-highest accuracy of 75% is achieved by MLP and Extra Tree classifier. The accuracy of KNN is 74.04%, while the accuracy of AdaBoost is 70.19%.

The Area Under the Curve (AUC) for different classifiers has been measured also. When the value of AUC for a certain classifier is 1.0 then the classifier is considered as a perfect classifier and if the value of AUC is 0.5, then the classifier is considered as a worthless classifier. Here the achieved AUC value for the Weighted Voting Classifier is 0.90 which has surpassed the AUC value of other base classifiers and the lowest AUC value was achieved by the Logistic Regression. The AUC value of Logisitic Regression is 0.76. The AUC values of SVM, KNN, Decision Tree, AdaBoost, MLP, and Extra Tree classifier are 0.81, 0.85, 0.87, 0.83, 0.81, and 0.85 respectively.

Additionally, this study has been performed by reducing the number of class labels also. In this task, the classes 'Excellent' & 'Good' have been categorized as 'Higher Order Grades' and 'Poor' & 'Fail' classes have been categorized as 'Lower Order Grades'. After that the performance of the proposed approach has been measured again. Table V represents the performance metrics using two class labels. From this table, it is found that the Weighted Voting Approach has gained the highest accuracy of 93.26%. Comparing Table IV and Table V, it can be observed that the accuracy has been significantly increased by reducing the number of class labels.

In both cases, the Weighted Voting Classifier has improved the accuracy. So, it can be confirmed that the Weighted Voting Classifier has overshadowed the other classifiers undoubtedly.

This study uses a dataset of 400 students of different courses and different departments. As the study has gathered the data from a variety of departments and with this dataset the proposed approach gained an accuracy of 81.73%, so, it can be assured that the proposed approach is reliable enough.

V. Conclusion

This study has used seven base classifiers to predict the students' final grades and then combined the outputs of the base classifiers using weighted voting approach. And from the observation, it can be confirmed that aggregating the base classifiers using the weighted voting approach has caused a rise in the accuracy. From the achieved AUC values it can be also stated that the Weighted Voting Classifier is almost the perfect classifier for classifying the accumulated dataset.

The limitation of this study is, it has not shown any comparison among the performance of this proposed approach and other approaches' performance, illustrating other study.

VI. Future Works

This work is performed by using the dataset of only one private university of Bangladesh. In future the dataset can be enlarged by collecting data from different private and public universities of Bangladesh to achieve better performance and better accuracy. Moreover, a comparative study between the proposed approach of this work and the approaches presented in other works can be performed in future.

Different studies show that by preprocessing the data using discretization method and oversampling techniques like Synthetic Minority Over-sampling Technique (SMOTE) can result in an increase of the accuracy. By using these approaches, preprocessing of the gathered dataset can be performed to get better performance.

References

[1] "Private University," Available online: http://www.ugc-universities.gov.bd/private-universities [Last Accessed 27 February 2020].

[2] S. K. Yadav and S. Pal, "Data mining: A prediction for performance improvement of engineering students using classification," arXiv preprint arXiv: 1203.3832, 2012.

[3] R. Kabra and R. Bichkar, "Performance prediction of engineering students using decision trees," International Journal of computer applications, vol. 36, no. 11, pp. 8–12, 2011.

[4] A. M. Shahiri, W. Husain et al., "A review on predicting student's performance using data mining techniques," Procedia Computer Science, vol. 72, pp. 414–422, 2015.

[5] B. Kapur, N. Ahluwalia, and R. Sathyaraj, "Comparative study on marks prediction using data mining and classification algorithms," International Journal of Advanced Research in Computer Science, vol. 8, no. 3, 2017.

[6] Z. Liu and X. Zhang, "Prediction and analysis for students' marks based on decision tree algorithm," In 2010 Third International Conference on Intelligent Networks and Intelligent Systems, pp. 338–341. IEEE, 2010.

[7] M. Sweeney, J. Lester, and H. Rangwala, "Next-term student grade prediction," In 2015 IEEE International Conference on Big Data (Big Data), pp. 970–975. IEEE, 2015.

[8] S. K. Yadav, B. Bharadwaj, and S. Pal, "Data mining applications: A comparative study for predicting student's performance," arXiv preprint arXiv: 1202.4815, 2012.

[9] B. Minaei-Bidgoli, D. A. Kashy, G.Kortemeyer, and W. F. Punch, "Predicting student performance: an application of data mining methods with an educational web-based system," In 33rd Annual Frontiers in Education, 2003. FIE 2003., vol. 1, pp. T2A–13. IEEE, 2003.

[10] Z. Iqbal, J. Qadir, A. N. Mian, and F. Kamiran, "Machine learning based student grade prediction: A case study," arXiv preprint arXiv: 1708.08744, 2017.

[11] W. L. Gorr, D. Nagin, and J. Szczypula, "Comparative study of artificial neural network and statistical models for predicting student grade point averages," International Journal of Forecasting, vol. 10, no. 1, pp. 17–34, 1994.

[12] Y. Meier, J. Xu, O. Atan, and M. Van der Schaar, "Predicting grades," IEEE Transactions on Signal Processing, vol. 64, no. 4, pp. 959–972, 2015.

[13] S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, "Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique," Decision Analytics, vol. 2, no. 1, p. 1, 2015.

[14] Z. Kovacic, "Early prediction of student success: Mining students' enrolment data," 2010.

[15] S. T. Hijazi and S. Naqvi, "Factors affecting students' performance," Bangladesh e-journal of Sociology, vol. 3, no. 1, 2006.

[16] Md. Jueal Mia, Al Amin Biswas, Abdus Sattar, Md. Tarek Habib, " Registration Status Prediction of Students Using Machine Learning in the Context of Private University of Bangladesh," International Journal of Innovative Technology and Exploring Engineering(IJITEE), vol. 9, no. 01, pp. 2594-2600, 2019.

[17] Al Amin. Biswas, Anup Majumder, Md. Jueal Mia, Itisha Nowrin, and Nadia Afrin Ritu, "Predicting the Enrollment and Dropout of Students in the Post-Graduation Degree using Machine Learning Classifier," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8 no. 11, pp. 3083-3088, September 2019.

[18] M. H. Bhuyan and S. S. A. Khan, "Motivating students in electrical circuit course," International Journal of Learning and Teaching, vol. 10, no. 2, pp. 137–147, 2018.

[19] R. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-level sentiment classification: An empirical comparison between svm and ann," Expert Systems with Applications, vol. 40, no. 2, pp. 621–633, 2013.

[20] K. Eashwar, R. Venkatesan, and D. Ganesh, "Student performance prediction using svm," International Journal of Mechanical Engineering and Technology, vol. 8, no. 11, pp. 649–662, 2017.

[21] M. Olgun, A. O. Onarcan, K. Özkan, Ş. Işik, O. Sezer, K. Özgişi, N. G. Ayter, and Z. B. Başçiftçi, M. Ardiç, and O. Koyuncu, "Wheat grain classification by using dense SIFT features with SVM classifier," Computers and Electronics in Agriculture, vol. 122, pp. 185-190, 2016

[22] A. Dutta, G. Bandopadhyay, and S. Sengupta, "Prediction of stock performance in indian stock market using logistic regression," International Journal of Business and Information, vol. 7, no. 1, 2012.

[23] Y. Y. Aung and M. M. Min, "Hybrid intrusion detection system using k-means and k-nearest neighbors algorithms," In 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), pp. 34–38. IEEE, 2018.

[24] S. K. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, and classifiaction," 1992.

[25] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Machine learning, vol. 63, no. 1, pp. 3–42, 2006.

[26] D. Tripathi, D. R. Edla, V. Kuppili, A. Bablani, and R. Dharavath, "Credit scoring model based on weighted voting and cluster based feature selection," Procedia computer science, vol. 132, pp. 22–31, 2018.

Vehicle Routing Optimization for Surplus Food in Nonprofit Organizations

Ahmad Alhindi[1], Abrar Alsaidi[2], Waleed Alasmary[3], Maazen Alsabaan[4]
Computer Science Department, College of Computer and Information Systems
Umm Al-Qura University, Saudi Arabia, Makkah[1,2]
Computer Engineering Department, College of Computer and Information Systems
Umm Al-Qura University, Saudi Arabia, Makkah[3]
Computer Engineering Department, Computer Science and Information College
King Saud University, Saudi Arabia, Riyadh[4]

***Abstract*—Non-profit organizations mitigate the problem of food insecurity by collecting surplus food from donors and delivering it to underprivileged people. In this paper, we focus on a single non-profit organization located in Makkah city (Saudi Arabia), referred to as Ekram. The current surplus food pickup/delivery and operational routing model of Ekram organization have several apparent deficiencies. First, we model the surplus pickup/delivery and operational routing model as a vehicle routing problem (VRP). Then, we optimize the pickup/delivery of different types of food groups through the different available routes. Finally, we utilize the formulated VRP problem by minimizing the total route distances. Our proposed model ensures reduction of the total time and effort necessary to send the collecting vehicles to the donors of surplus food.**

Keywords*—Non-profit organization; vehicle routing problem; donor; surplus food; decision support

I. Introduction

Food waste is a globally critical issue and requires collective responsibility to be appropriately addressed. Fortunately, there exist several dedicated non-profit organizations to support food security by providing quality food surplus services [1]. The well-known charities that distribute food to underprivileged people in North America are Action Against Hunger, Feeding America and Food for the Poor. These organizations rely heavily on people donations. The donation-driven food delivery system is different from commercial supply chains because the commercial version rewards fast and cost-effective delivery as they generate extra profits [1]. While the food redistribution in non-profit organizations seeks to move the surplus food obtained through donations from the donors to the underprivileged people efficiently, the overarching objective is, therefore, nonprofit organization [2].

The St. Mary's Food Bank Distribution Centre adopted the heuristic concept that ensured efficiency in the track distribution system over the network. The organization relies on the services of about 330 agencies to enhance the efficient delivery of food products [3]. The VRP of the organization relies on the structured location of the agencies for better service delivery at minimal costs. The VRP models adopted by the food service organizations serve in the generation of minimum routes that serve most customers [4]. The problem of late collection of donated food staff from the donor and loading of tracks from the food banks severely affects food distribution since any delay accelerate the extent of wastage. Therefore, the presence of close monitoring and delivery of products significantly eliminate the probable long-term wastage in the service industry [5].

The food donations are different in terms of their quantity and frequency [1]. Hence, a comprehensive system is required to coordinate the collection and distribution of donations in an efficient mechanism. The mechanism has to take into consideration the following parameters: (1) avoid wasting the food, (2) reduce the delivery time, and (3) use the minimum number of drivers. In this paper, we model surplus food pickup/delivery operations of a non-profit organization in Makkah Province (Saudi Arabia) as an optimization problem, namely, a vehicle routing problem (VRP) that arise in many practical situations, e.g, pickup and delivery of goods to customers. The contributions of this proposed solution are as follows.

- We propose a novel routing model specifically designed for food surplus pick up/delivery systems to optimize route selection and time delivery.
- Improve the performance of the system by effective fleet management of transport.
- Improve vehicle utilization (one vehicle can pick up more than one package in a single route).

The rest of this paper is organized as follows. Section II provides a review of related works. Section III discusses the features of the Ekram organization. The proposed solution that will be used to solve the research problem is given in Section IV. Section V describes the proposed VRP model. Section VI describes the simulation study. Finally, the conclusion is given in Section VII.

II. Related Works

The implication of the Vehicle Routing Problem (VRP) uses an integrated optimization program to facilitate the logistics for the fleets of vehicles in managing the distribution systems. The programs facilitate the management of resources for efficient operations in complex networks. The efficiency with which the organizations manage the VRP define the sustainability of the enterprise in the non-profit entity platform [6].

A. VRP in Food Delivery Organizations

The Feeding Americans organization coordinates over 200 food banks and 60000 distributors [7]. It depends on a well-coordinated network of the VRP. The organization uses the heuristic solution approach to create a balance between the routs and the duration spent to reach the community clients [7].

Another example is the OzHarvest which is a food bank in Australia [8]. The objective of this organization is collecting and redistributing foodstuffs to the welfare agencies supporting in the region based on the food recovery techniques to minimize wastes. The primary donations originate from restaurants and large food store outlets in the area. The advantages of this organization can be formulated in three points:

1) The focus on product specialization limits the extent of trash within the organization [8].
2) The organization implements successful training to the drivers to be able to check products at the collection point and enact reroute as deemed relevant based on the shelf life of the product donated.
3) The organization implements successful training to the drivers to be able to check products at the collection point and enact reroute as deemed relevant based on the shelf life of the product donated.

The use of VRP forecasting models facilitates efficiency in the organization through the implementation of overrun and underrun based on the peak threshold in the food supply platform [8].

Also, the Good Shepherd Food Bank is an example that should be mentioned in this article. This organization is located in Maine City [7]. It relies on food donations from the emergency food rescue team, the supermarkets, and the soup kitchen. In order to improve both the design and operations, the organization tends to collaborates with other food agencies. Such organizations encountered many constraints like the distance between the donation, consumer, and warehouse locations in addition to limited refrigeration facilities. Therefore, the focus on distribution network topology improves the level of VRP implementation in the organization [7]. The VRP design focuses on the population density of the consumer to facilitate re-routing of the delivery to accommodate efficiency in terms of management and operation of the agency outlets.

In [9], the researcher talks about the central depot of food supplies to other branches of Greater Bangkok and the losses experienced due to the perishability of food supplies. So, the central depot implements cooling trucks designed by routing and scheduling program in VRP. This minimized the loss of perishable goods such as vegetables. This goes a long way in saving food security in the city of Great Bangkok. It also creates the reliability of the central depot for the delivery of food supply.

A green logistic case study is provided in [10]. The researchers talk about Eroski, which is a Spanish supermarket chain, and highlights the challenges that the company suffers from late due to late deliveries. This study considers environmental pollution caused by the emission of several vehicles during deliveries of food products to the market [10].The delivery department at Eroski Company employed the VRP through the following methods:

1) It enabled drivers to deliver food supplies in time since VRP allows them to operate on different routes compared to the previous trucks where they could only access few routes.
2) It enabled the company to make fast deliveries.
3) It minimized the environmental pollution by taking the benefit of the fact that some vehicles can collect and deliver waste of food on various routes [10].

The last article that will be mentioned here is about Optimization of Vehicle Routing for Smart City. This article investigates the issue that supermarkets suffer from lack of availability of food products in the store due to lack of sufficient delivery in Casablanca [11]. The supermarkets in Casablanca addressed the issue of the unavailability of the on-shelf product by integrating geographic information systems (GIS) with the VRP. The GIS was used to incorporate actual vehicle travel data and then choose the optimal structure to satisfy customers in order to minimize late deliveries and reduce vehicle congestion.

III. Ekram Organization

Ekram is a Saudi non-profit organization that officially declared by the Ministry of Labor and Social Development in 2007. This organization is located in the city of Makkah on the western side of Saudi Arabia. Ekram is a food rescue organization that plays a crucial role in alleviating hunger by preserving the food resources in the society by providing the complete process of picking up surplus food from donors up to delivering it the underprivileged people. The primary objective of Ekram is to create awareness on the importance of saving surplus food and not throwing it, through the provisioning process of it to the needy based on health requirements and distributed both meals and water for people in Ramadan and Hajj season. This organization's mission is saving the food by packaging the surplus food and delivering it to the underprivileged people in a hygienic condition. The organization's vision is to collect surplus, rapidly distributed using the best food safety standards (see http://www.ekram.org.sa).

Furthermore, the organization aims to manage the efficient communication and coordination of the collection of surplus food from donors and to distribute it to the underprivileged people. Ekram will also promote access to quality healthy food to the needy in society as a basic form of human wants. Considering that this organization's primary aim is to redistribute food resources in society, it is equipped with 40 vehicles of different capacities. The organization has also partnered with donors such as hotels as the sources of surplus food.

Ekram seeks to achieve the Saudi Vision 2030 through

- Reduce wastage of food by following the latest international standards and experiences.
- It is motivating people to volunteer.

One case that should be consider is when having more than a donor. In this case, several points should be taken in consideration, such as the fact that the food is going to be collected by different vehicles or depending on the availability of the automobiles (bear in mind here that there is little consideration of the route taken), the amount of the foods, the capacity of the vehicles (one vehicle delivered the food as a parts in several rounds from source to the destination), the proximity of multiple donors. Moreover, drivers will be required to decide depending on the kind and quality of surplus food collected from donors, on whether to distribute it to underprivileged people immediately or to deliver it to the center for sorting and packaging before entering the distribution phase. Efficient decision making by the drivers is based on several factors such as time, vehicle capabilities, and routing. Due to the lack of a well-defined vehicle planning system, the routing process in Ekram is often randomly.

The primary challenge facing Ekram is saving both time and efforts (in terms of sending vehicles to donors for collection of the surplus foods and how the requests can be managed).

IV. Proposed Solution for Ekram: Vehicle Routing Problem

The vehicle routing problem (VRP), also called route optimization, is described as a fleet of vehicles with different capacities, and a typical depot with several customers' demands to locate a set of routes that will offer minimum cost and serve all requests [12].

In the VRP, all the itineraries commence and terminate at the depot and are designed in such a way that each request is served only once by a single-vehicle [13]. The VRP is increasing the transport efficiency of a fleet of vehicles by finding the shortest route for each vehicle and considering real-time updates as the statuses of items change.

Theoretically, the VRP is a combinatorial optimization problem that may apply the integer programming technique to solve. The method seeks to determine the optimal set of routes for a fleet of the vehicle such as in the case of Ekram where the vehicles are expected to collect the surplus food and deliver in a depot where it can be distributed to the underprivileged people. The primary aim of the conventional vehicle routing techniques is to minimize the cost of a route. In the VRP employs several vehicles and set of routes; thus, the VRP is constrained by two vital elements, which are

1) Time constraints: every location in the route must visit within a particular period.
2) Capacity constraint: considering that the vehicles are required to collect items from each location, but a maximum capacity limits them.

The VRP will work on minimizing the length of the longest route for all vehicles. This assists in attaining the aim of VRP by completing all the deliveries within the shortest time possible.

V. Proposed VRP Model

The VRP aims to reduce both the time and effort involved in sending the vehicle to the donor for the collection of surplus food and the management of this problem. The VRP will be used to determine the low cost shortest path to service the charity demands.

In the case of Ekram is define on graph $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, where $\mathcal{N}$ is the sets of nodes (destinations) and the depot $\mathcal{N} = \{n_i, \cdots, n_m\}$, at different locations. Every pair of locations are denoted by (i,j), $\mathcal{A}$ set of arcs such that $\mathcal{A} = \{(i,j) : i,j \in \mathcal{N}, i \neq j\}$, $\mathcal{C}$ represents the set of donor associated with Ekram organization, $\mathcal{V} = \{v_i, \cdots, v_m\}$ is the sets of vehicles, y_{ik}, y_{ijk} and z_{ijk} are the binary variables, x_{ijk} is the continuous variable, L is the vehicle capacity, i and j are nodes, k is the vehicles that have visited a node, and d_i is the demand at the node.

We define binary decision variables z_{ijk} equal to 1 if and only if in optimal solution vehicle k visits customer j after customer i, and y_{ik} equal to 1 if and only if vertex i is served by vehicle k. The problem is formulated as follows:

$$\min_{z_{ijk}} \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{A}} c_{ij} z_{ijk} \quad k \in \mathcal{V} \tag{1}$$

subject to the following terms:

$$\sum_{k \in \mathcal{V}} y_{ik} \geq 1 \qquad i \in \mathcal{C} \qquad \text{(Assignment)} \tag{2}$$

$$\sum_{(i,j) \in \mathcal{A}} z_{ij} = y_{ijk} \quad i \in \mathcal{N}, k \in \mathcal{V} \qquad \text{(LeaveNode)} \tag{3}$$

$$\sum_{(i,j) \in \mathcal{A}} z_{ij} = y_{ijk} \quad i \in \mathcal{N}, k \in \mathcal{V} \qquad \text{(EntryNode)} \tag{4}$$

$$\sum_{(i,j) \in \mathcal{A}} x_{ijk} - \sum_{(i,j) \in \mathcal{A}} x_{ij} = d_i y_{ik} \quad i \in \mathcal{C}, k \in \mathcal{V} \qquad \text{(FlowBalance)} \tag{5}$$

Determining the vehicle capacity.

$$x_{ijk} < L z_{(ijk)} \quad (i,j) \in \mathcal{A}, k \in \mathcal{V} \qquad \text{(VehicleCapacity)} \tag{6}$$

$$y_{1k} = 1 \quad i \in \mathcal{C}, k \in \mathcal{V} \qquad \text{(Depot)} \tag{7}$$

The objective function aims to minimize the total traveled distance. Constraint (2) ensuring that each customer is served by at least one vehicle. The constraints (3) and (4) describing the arc within which the vehicle will leave or visit node i. The leave and entry node enforce the following condition: if node i is visited by vehicle k then vehicle k will use one arc entering the node and then another arc that leaves node i. Additionally, if node i is not visited by vehicle k, then no arc entering or leaving node i should be used by vehicle k. Moreover, the constraint (5) describes the flow conservation at the nodes for each vehicle and constraint (6) ensure that no vehicle can be overloaded, the quantity of the product in each vehicle must always be less or equal to the vehicle capacity L. Finally, constraint (7) show when the vehicle is at the depot.

VI. Simulation Study

A. Simulation Setup

We performed a simulation study of an organization similar to Ekram, and we studied the organization operating for a whole month. All the simulation experiments have been carried out on MacBook Pro (Intel Core i5 1.4GHz CPU and 8GB RAM). The programming language is Java.

First, we are assuming that there are 20 stores and 30 drivers or vehicles. Moreover, each store and driver has a unique ID number. Second, vehicles and drivers are uniformly distributed over an area of (10km x 10km).

In our simulation, in every working hour, the organization takes the order from the store, which includes the store ID and the amount of surplus food. Then, the program selects the vehicle in order to send it to the store. Next, the program computes the distance between the selected vehicle and the store in order to divide the distance by average vehicle speed and computes the time for each vehicle. Finally, the program calculates the amount of food received by the end of each day.

Here, we assume that there are 20 orders for the next hour so that drivers will be assigned randomly to these orders. Fig. 2 shows the output that will be achieved.

After that we assume the same number of orders on the first case but with applying the VRP (involves the choice of short-distance routes), before assigning the driver to the order. Fig. 3 shows the output.

In this study, we will measure performance based on two of the performance metrics

1) Travel time: the time spent by the vehicle to reach its destination.
2) Travel distance: the route distance from the vehicle location to the destination.

We assume that each pixel is equal to two meters and the average speed of vehicle is (80 km/h). So, the travel time will be calculated based on the following formula:

$$Travel\ time = distance \div speed \tag{8}$$

Moreover, the travel distance will be calculated based on the following formula:

$$Travel\ distance = speed \times time \tag{9}$$

B. Simulation Results and Discussion

In this section, we describe the results obtained by our simulation. The performance measures used to evaluate the results are the total distance driven (travel distance) and total travel time.

Fig. 1 shows a graph/map of the stores and drivers locations. The black point represents the stores while the green points represent the drivers Moreover, Fig. 2 shows the drawbacks of the random selection, which can be summarized in two points. The first one is that some drivers travel more than others and the second one is that some drivers assigned to stores far to them.

In the simulation study above, when applying the VRP on it, the program assigned a driver based on its location relative to the store, which will lead to using the shortest path, as shown in Fig. 3.

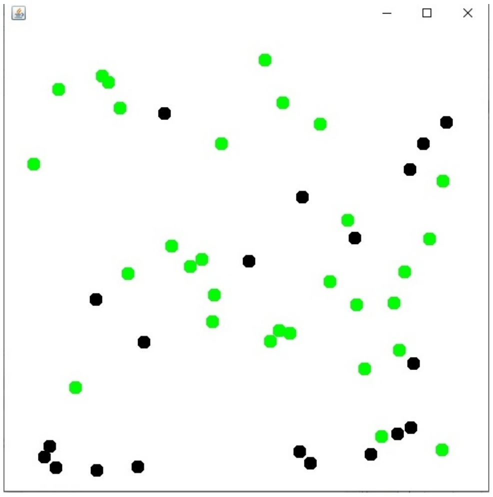


Fig. 1. An illustration of the location of the pick up/delivery points and the driver points on the hypothetical map. A one hour simulation scenario. The black point represents the stores while the green points represent the drivers.

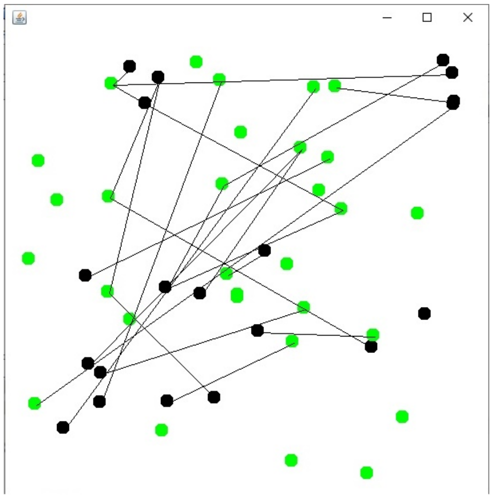


Fig. 2. An illustration of the simulation study to demonstrate the **random mapping** (i.e., creating a random graph/ assignment of drivers to pickup/delivery locations randomly). A one hour simulation scenario.

After implementing the simulation study for a complete month, the proposed VRP model (PRO) shows a smaller total for both travel time and distance when comparing to the travel time and distance obtained in the random model (RAND). Table I shows the total of travel time, travel distance, and number of the vehicle used at two models. At the bottom of the table, we add a row to show the utilization of the VRP model over the random model. The utilization is calculated in terms of percentage in terms of each of the above metrics as follows:

$$\begin{aligned}\text{Utilization}(\%) = [(&\text{RAND}_{\text{metric value}} - \text{PRO}_{\text{metric value}}) \\ &\div \text{RAND}_{\text{metric value}}] \times 100\end{aligned} \tag{10}$$

Table I shows the improvement in travel time with 59.7% and travel distance by 46.5% by using the VRP model comparing to the random model. This is because the VRP

TABLE I. Comparison of travel time (hour), travel distance (meter) and number of used vehicles with Random and VRP models. A one month simulation scenario.

	Travel time (hour)	Travel distance (meter)	Number of vehicles
Random model	467	64989494.0671	30
VRP model	188	34746409.7178	17
Utilization %	59.7%	46.5%	43.3%

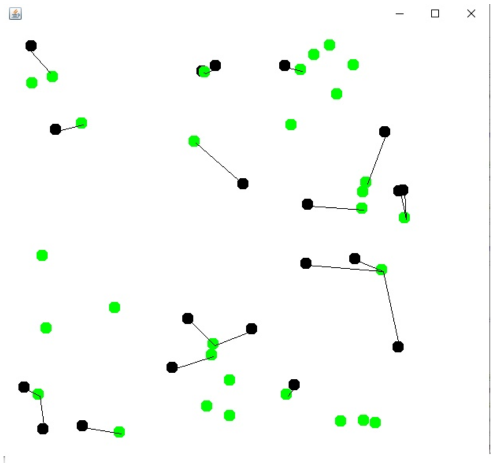


Fig. 3. An illustration of the simulation study to demonstrate the **proposed VRP mapping** with only 20 resulting links compared to Fig. 2. A one hour simulation scenario.

model searches for routes with the least distance before assign vehicles, which reduces the time is taken and enables one vehicle to service more requests. Thus, reduce the total number of vehicles used to satisfy the demands from 30 vehicles of the random model, it reduced to 17 vehicles of the VRP model. From that result, the reduction 43.3% in the VRP model.

For Ekram, VRP would be an excellent way to plan the collection and distribution of food effectively. However, the solution would have to be customized according to the needs and specific activities of the organization. For instance, a system for communicating the status of donors would need to implement, such that both the driver and the technical support team at the center can see whenever a new donation is available. Additionally, a database of common way points would also be necessary to increase the speed of the Route Optimizer. Based on the position of each of the 40 vehicles, the system can then suggest the most appropriate driver collect the new donation and determine whether it is more efficient to return it to the center or distribute it immediately.

Additionally, it would be necessary to incorporate information on the capacity of each vehicle, as well as any restrictions on the type of food it can carry. In such a way, the algorithm can compare the type and amount of food donated and determine the vehicle that should go and collect it. In the case of multiple donors, a vehicle can collect food at more than one point if it has enough capacity, and the most optimum route is selected for each automobile to avoid unnecessarily long trips or movement of near-empty vehicles.

When selecting the optimal path by Ekram, which is the shortest path with fewer efforts used to collect the food the following will be the outcomes:

- Enhance service quality by minimizing operational costs.
- Minimize the total distance of route (by determining the best path to reach the destination).
- Save both time and money used in the collection of food from the donor by using the optimal route.
- Improve the speed of movement of the vehicles between the depot and the food collection points.
- Increase the satisfaction of the needy people as well as the donors because both delivering and collecting waste of food will be done based on the scheduled plans.
- Minimize the number of required vehicles to meet the total demand.

By accounting for all ongoing and completed orders and employing advanced methods to analyze the route in real-time. VRP can tremendously increase transport efficiency. Besides that, it enhanced the quality and reliability of vehicle routing decisions by mitigating subjectivity in the decision-making process.

VII. Conclusion

The need to collect and deliver multiple items in the shortest possible time while maintaining operational efficiency is a significant challenge for food delivery organizations. Ekram, which is a Saudi non-profit food saving organization, has a daily task of collecting and re-distributing surplus food in the city of Makkah. This daily routine of Ekram lacks a systematic decision support and efficient vehicle planning. One effective way to address this challenge is to utilize the vehicle routing problem (VRP) to better manage and determine the optimal set of routes a fleet of vehicles should take. In this paper, we demonstrate that the utilization of the VRP into the surplus food pickup/delivery model of Ekram organization results in an increased efficiency, reduced cost, and enhanced better management of the surplus food. Our simulation study demonstrated that our proposed model significantly outperforms the random pickup/delivery model in terms of the travel time and distance.

References

[1] L. B. Davis, S. X. Jiang, S. D. Morgan, I. A. Nuamah, and J. R. Terry, "Analysis and prediction of food donation behavior for a domestic hunger relief organization," *International Journal of Production Economics*, vol. 182, pp. 26–37, 2016.

[2] D. J. Nair, T. H. Rashidi, and V. V. Dixit, "Estimating surplus food supply for food rescue and delivery operations," *Socio-Economic Planning Sciences*, vol. 57, pp. 73–83, 2017.

[3] X. Li, *Capacitated Vehicle Routing Problem with Time Windows: A Case Study on Pickup of Dietary Products in Nonprofit Organization*. Arizona State University, 2015.

[4] N. Labadie, C. Prins, and C. Prodhon, "General presentation of vehicle routing problems," *Metaheuristics for Vehicle Routing Problems*, vol. 3, pp. 1–14, 2016.

[5] D. Goeke, R. Roberti, and M. Schneider, "Exact and heuristic solution of the consistent vehicle-routing problem," *Transportation Science*, vol. 53, no. 4, pp. 1023–1042, 2019.

[6] G. Erdoğan, F. McLeod, T. Cherrett, and T. Bektaş, "Matheuristics for solving a multi-attribute collection problem for a charity organisation," *Journal of the Operational Research Society*, vol. 66, no. 2, pp. 177–190, 2015.

[7] L. B. Davis, I. Sengul, J. S. Ivy, L. G. Brock III, and L. Miles, "Scheduling food bank collections and deliveries to ensure food safety and improve access," *Socio-Economic Planning Sciences*, vol. 48, no. 3, pp. 175–188, 2014.

[8] D. J. Nair, H. Grzybowska, D. Rey, and V. Dixit, "Food rescue and delivery: Heuristic algorithm for periodic unpaired pickup and delivery vehicle routing problem," *Transportation Research Record*, vol. 2548, no. 1, pp. 81–89, 2016.

[9] K. Panapinun and P. Charnsethikul, "Vehicle routing and scheduling problems: a case study of food distribution in greater bangkok," *Bangkok: Kasetsart University*, 2005.

[10] S. Ubeda, F. J. Arcelus, and J. Faulin, "Green logistics at eroski: A case study," *International Journal of Production Economics*, vol. 131, no. 1, pp. 44–51, 2011.

[11] L. Safia, B. Jamal, A. Mustapha, M. Salma, and A. H. Sabry, "Optimization of vehicle routing for smart city: Real case study in casablanca," *Smart Application and Data Analysis for Smart Cities (SADASC'18)*, 2018.

[12] H. Kurnia, E. G. Wahyuni, E. C. Pembrani, S. T. Gardini, and S. K. Aditya, "Vehicle routing problem using genetic algorithm with multi compartment on vegetable distribution," in *IOP Conference Series: Materials Science and Engineering*, vol. 325, no. 1. IOP Publishing, 2018, p. 012012.

[13] A. K. M. Masum, M. Shahjalal, F. Faruque, and I. Sarker, "Solving the vehicle routing problem using genetic algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 7, pp. 126–131, 2011.

Development of an Interactive Tool based on Combining Graph Heuristic with Local Search for Examination Timetable Problem

Ashis Kumar Mandal
Faculty of Software and Information Science
Iwate Prefectural University
Iwate, Japan

***Abstract*—Every university faces a lot of challenges to solve the examination timetabling problem because the problem is NP-hard and contains numerous institutional constraints. Although several attempts have been taken to address the issue, there are scarcities of interactive and automated tools in this domain that can schedule exams effectively by considering institutional resources, different constraints, and student enrolment in courses. This paper presents the development of a system as a graphical and interactive tool for examination timetabling problem. To develop the system, combining graph coloring heuristic and local search meta-heuristic algorithms are employed. The graph heuristic ordering is incorporated for constructing initial solution(s), whereas the local search meta-heuristic algorithms are used to produce quality exam timetables. Different constraints and objective functions from ITC2007 exam competition rules are adopted, as it is a complex real word exam timetabling problem. Finally, the system is tested on the ITC2007 exam benchmark dataset, and test results are presented. The main aspect of the system is to deliver an easy-to-handle tool that can generate quality timetables based on institutional demands and smoothly manage several key components. These components are collecting data associated with the enrolment of students in exams, defining hard and soft constraints, and allocating times and resources. Overall, this software can be used as a commercial scheduler in order to provide institutions with automated, accurate, and quick exam timetable.**

***Keywords*—Examination timetable; graph heuristic; local search meta-heuristic; ITC2007 exam dataset; interactive tool; NP-hard problem**

I. Introduction

Solving the examination timetable of an academic institution induces lots of complicacies due to the intractable nature of the problem. That is, managing different constraints and producing a quality exam timetable, which meets the demand of the institution, often computationally expensive and laborious task. The procedure is so complicated that human schedulers also struggle to produce even a simple feasible solution. Some viable approaches are operations research (OR) and artificial intelligence (AI) techniques that handle the problem often by mathematical programming as well as different heuristics, meta-heuristic, and hyper-heuristic algorithms [1]. Some of these procedures are constraint programming [2], integer programming [3], graph heuristics [4], great deluge algorithm [5], hill-climbing [6], tabu search [7], simulated annealing [8], genetic algorithm [9], particle swarm optimization [10], artificial bee colony algorithm [11] and memetic algorithm [12].

It is frequently observed that academic institutions rely on traditional manual approaches, which take considerable time on managing the vast amount of student registration data, conflicting exams, as well as the violation of constraints for producing a feasible timetable. Although numerous approaches have been proposed in the literature for creating quality timetables, the work on interactive software for timetabling problems is limited. In the examination timetabling survey, Qu et al. [13] emphasize reducing the gap between research and practice and highlight the importance of automatic interactive examination timetable tools for reducing the significant workload of timetabling staff. There are some software solutions proposed for university timetabling problems so that the user can easily interact with the timetable generations. Piechowiak and Kolski [14] developed interactive tools for supporting University of Valenciennes and Hainaut-cambresis (UVHC) university timetabling. The aim was to develop an open, generic tool that employs distributed architecture for cooperative scheduling and supports users to monitor modeling of time, resource, university activities, and constraints for producing a quick feasible solution. Thomas et al. [15] proposed a visual interface tool that aids users to quickly find a bottleneck situation and guide the scheduling system towards a feasible solution. Ayob et al. [16] proposed an intelligent examination timetabling software. The aim was to develop an intelligent commercial scheduler that can replace human decision-makers as well as produce high-quality solutions for University Kebangsaan Malaysia (UKM) examination timetabling problem. Chunbao and Nu [17] developed an efficient exam scheduling system (IIEESS v1.0) that avoids the traditional direct-clash-checking approach and schedules a large number of exam papers within a few minutes. Another recent software solution is solving the University of Toulouse examination timetabling problem using integer linear programming [18]. The authors claimed that the tool can produce quality solutions automatically and give some flexibility to choose input data and constraints.

Although the above tools are viable in producing quality exam timetabling, most of them emphasize solely on automation rather than interaction with users. Besides, interactive tools for generating student conflicts performed by open registration, simultaneous execution of different optimization processes, and handling more complex real-world problems like ITC2007 exam dataset are some issues that have been less highlighted.

This paper has proposed an interactive tool for addressing the examination timetabling issue for universities. The system initially produces conflicting exams automatically from course enrollment data. Then it facilitates in managing various hard and soft constraints and allocating times and resources. Based on that configuration, the initial solution module, which uses a saturation degree graph heuristic (SD), produces a feasible solution. Users can monitor the solution quality with different configurations and even further improve the solution vector with employing an improvement module. This module contains three local search algorithms: Great Deluge Algorithm (GDA), Simulated Annealing (SA), and Late Acceptance Hill Climbing (LAHC). A user can select different algorithms, tune parameters, and inspect the progression of the solution. In addition, another facility is the execution of concurrent run with selected local search algorithms. The improvement phase finally produces a solution vector along with a quality metric within specific stopping criteria. The proposed system has been tested successfully using ITC2007 exam benchmark dataset, which covers the majority of the hard and soft constraints of many real exam timetables. The goal in this paper is to provide an effective interactive examination timetable software such that it can generate computationally inexpensive quality timetable solutions and reduce both context-dependencies and involvement of human expertise as much as possible.

The rest of this paper is organized as follows: Section II describes the examination timetabling problem formulation. Section III highlights the algorithms employed for building the system. The proposed system architecture and software component for addressing examination timetabling problems have been presented in Section IV. Section V presents simulation results and discussion. Finally, some conclusions are drawn in Section VI.

II. Problem Description and Formulations

An examination timetabling is a scheduling problem where a set of examinations is allocated into a limited number of time slots and rooms subject to a set of constraints. Generally, two different types of constraints encompassing hard constraints, and soft constraints must be addressed. Satisfying all hard constraints leads to a feasible solution, whereas the minimization of soft constants results in a quality solution. Frequently these soft constraints are associated with objective functions. All hard and soft constraints vary from one institution to another based on institutional requirements and resources.

Examination timetable problems can be categorized as capacitated and un-capacitated problems. In an un-capacitated branch, room capacity is not considered. In a capacitated variant, however, room capacity is considered as a hard constraint. For instance, Toronto dataset is a un-capacitated problem, whereas the Second International Timetabling Competition (ITC2007) exam dataset is a capacitated problem [13].

In this section, ITC2007 exam dataset is described here because it is the most recent real-world examination timetabling problem, which has lots of hard and soft constraints. Besides, the proposed system has been developed with ITC2007 exam timetabling problem in mind. ITC2007 exam dataset consists of eight problem instances. The comprehensive characteristics such as number of students, number of exams, number of slots, number of rooms, period hard constraints, room hard constraints, and conflict density are presented in Table I.

All hard and soft constraints involved with the examination timetabling problem reported in ITC2007 exam dataset are given below:

Hard Constraints

- H1. Any student cannot sit more than one exam at the same time.
- H2. The exam capacity should not exceed room capacity.
- H3. The exam length should not violate the period length.
- H4. Three ordering of exams must be respected.
  - Precedences: exam i will be scheduled before exam j.
  - Exclusions: exam i and exam j must not be scheduled at the same period.
  - Coincidences: exam i and exam j must be scheduled at the same period.
- H5. Room exclusiveness must be maintained. For example, an exam i must take place only in room number 206.

Soft Constraints

- S1. Two exams in a row(C_s^{2R}): Avoid the number of occasions where a student sits consecutive exams on the same day.
- S2. Two exams in a day (C_s^{2D}): Avoid the number of occasions where a student sits two exams in a day. Note that when exams are one after another, this is counted as Two Exams in a Row for avoiding duplication.
- S3. Spreading of exams (C_s^{PS}): Exams should be spread as evenly as possible over time periods.
- S4. Mixed duration (C^{NMD}): Avoid the number of occasions where exams with different durations are scheduled into the same room.
- S5. Scheduling of larger exams (C^{FL}): Avoid the number of occasions where the largest exams are assigned later in the timetable.
- S6. Room penalty:(C^{R}) Avoid the number of occasions where certain rooms with an associated penalty are used for scheduling.
- S7. Period penalty (C^{P}): Avoid the number of occasions where certain periods with an associated penalty are used for scheduling.

The objective function is formularized as in Eq. 1. It attempts to minimize the violation of soft constraints (penalty) as much as possible for producing good quality solutions without violating the hard constraints.

TABLE I: Features of ICT2007 exam dataset

Instances	**No. of students**	**No. of exams**	**No. of slots**	**No. of rooms**	**Period hard constraints**	**Room hard constraints**	**conflict density**
Exam_1	7,891	607	54	7	12	0	5.05%
Exam_2	12,743	870	40	49	12	2	1.17%
Exam_3	16,439	934	36	48	170	15	2.62%
Exam_4	5,045	273	21	1	40	0	15.00%
Exam_5	9,253	1018	42	3	27	0	0.87%
Exam_6	7,909	242	16	8	23	0	6.16%
Exam_7	14,676	1096	80	15	28	0	1.93%
Exam_8	7,718	598	80	8	20	1	4.55%

$$min \sum_{s \in S}(W^{2R}C_s^{2R} + W^{2D}C_s^{2D} + W^{PS}C_s^{PS}) + \\ W^{NMD}C^{NMD} + W^{FL}C^{FL} + C^R + C^P \qquad (1)$$

In this equation, W (with different subscriptions) stands for the related weight for each of the soft constraints, and S indicates a set of students. Table II shows weights of ITC2007 exam dataset. Note that associated weights are not included in C^P and C^R in the equation as these associated weights are already added in the definition. Explaining all constraints, instances, mathematical models of the ITC2007 exam tracks are so wordy that details explanation will be found in [19] and the website at http://www.cs.qub.ac.uk/itc2007.

TABLE II: Weights of ITC2007 exam dataset

Instances	W^{2D}	W^{2R}	W^{PS}	W^{NMD}	W^{FL}
Exam_1	5	7	5	10	5
Exam_2	5	15	1	25	5
Exam_3	10	15	4	20	10
Exam_4	5	9	2	10	5
Exam_5	15	40	5	0	10
Exam_6	5	20	20	25	15
Exam_7	5	25	10	15	10
Exam_8	0	150	15	25	5

III. Overview of Algorithms

A. Graph Heuristics

A simple Examination timetabling can be represented as a graph coloring problem. It is an undirected graph comprising a set of n vertices and a set of edges E, with vertices indicating exams while exams with common students indicate an edge. For example, If Exam_1 and Exam_2 have a common student, there will be an edge E. There are a predefined limited number of colors that signify time slots. The exams have to be assigned into time slots (i.e., coloring the graph) in such a way that no exams with a common student have the same timeslots (i.e., color). Graph heuristics are based on ordering strategies where examination with most difficulty is chosen for scheduling first so that finally, a feasible solution can be obtained. Various graph heuristic techniques measure examination difficulty. These are the largest degree (LD), largest weighted degree (LWD), Largest enrolment degree (LE), and saturation degree (SD) [20], [21]. The heuristics are described as follows:

- Largest degree (LD): This technique orders the exams based on the largest number of conflicting examinations.
- Largest weighted degree (LWD): This heuristic is similar to the largest degree except the exams are ordered based on the number of students in conflict.
- Largest enrolment (LE): The exams are ordered based on the number of registered students in the exams.
- Saturation degree (SD): The exams are ordered based on the number of remaining timeslots available; exams with the least number of available timeslots in the timetable are given priority to be scheduled first. SD is a dynamic heuristic where the ordering of exams is updated as the exams being scheduled.

B. Local Search Approaches

Graph heuristics can generate an initial solution that is not optimum enough to consider a quality timetable. Hence local search meta-heuristics are frequently used to reduce the soft constraint violations as much as possible to get a quality solution from the initial solution. In this paper, three meta-heuristics are presented as they have been used for the proposed systems.

1) Late acceptance hill-climbing: Burke and Bykov [22] proposed late acceptance hill-climbing(LAHC) for escaping local optima produced by a greedy hill-climbing approach. In greedy hill-climbing, the candidate solution is compared with the immediate current one, but LAHC uses a delay comparison mechanism where the candidate solution is compared with the solution of several iterations earlier. The algorithm starts with an initial feasible solution, and a new candidate solution is checked for acceptance in each iteration. A list of a specific length is used for memorizing the previous values of the current cost function used for acceptance criteria. Each time the candidate solution is compared with the last value of the list, and if better, it is accepted. When the acceptance procedure activates, the new cost is added at the beginning of the list, and the last element is deleted. The procedure is performed base on $v = ImodL$ formula, where L is the length of the frame, I is the i^{th} iteration, and v is the position.

2) Simulated annealing: Simulated annealing (SA) is a local search meta-heuristic technique based on a physical annealing process that probabilistically accepts some worst solutions to escape from the local optimum [23]. SA starts with a randomly generated initial solution, and in each iteration, it tries to improve the solution quality. If the neighboring solution is better than or equal to the current solution, it is replaced with the current one. Otherwise, acceptance of neighboring solution is decided on a probability function $exp(-\frac{f(s^*)-f(s)}{T})$, where $f(s^*)$ is a neighboring solution, $f(s)$ is the current solution, and T is a parameter known as temperature. Initially, the algorithm starts with a high T and periodically decreases the value using a cooling schedule until the temperature is zero or any terminal condition.

3) Great deluge algorithm: Great deluge (GDA) algorithm was proposed by Dueck [24]. The inspiration of this algorithm originated from the behavior in which a hill climber seeks a higher place to avoid the rising water level during the deluge. Like SA, this algorithm devises a mechanism to avoid local optima by accepting the worst solutions. SA uses a probabilistic function for accepting the worst solutions, whereas GDA uses a more deterministic approach for that purpose. It is also found that GDA depends less on parameter tuning compared to SA. The only parameter in the GDA algorithm is the decay rate, which is used for controlling the boundary or acceptance level. In the minimization problem, the initial boundary level (water level) usually starts with an initial solution. During the search, a new candidate solution is accepted if it is better than or equal to the current solution. However, the solution worse than the current one will be accepted if the quality of the candidate solution is less than or equal to a predefined boundary level B. The boundary level then is lowered by subtracting a parameter called decay rate (ΔB). This parameter is vital because the speed of the search depends on the decay rate.

IV. System Architecture and Software Component

Fig. 1 depicts the overall system architecture of the examination timetabling scheduler. It consists of four different components: planning module, scheduling module, reporting module, and user interface.

A. Planning Module

Planning module deals with all the input data required for generating the solution. Here following steps are performed.

1) Constraints manager: Constraints manager handles all the hard and soft constraints associated with exam timetabling. Constraints can be modified according to user choice. Besides, another important function handled by this module is to compute the penalty cost of an exam solution. Here the objective function is employed for generating the penalty value of a given solution. This objective function can be predefined or it might be varied from one problem to another.

2) Exam conflict matrix: After all the necessary data are loaded into the system and constraints are defined, analysis of exam confliction is performed using a data structure named conflict matrix. The examination conflict matrix is a square matrix of dimension equal to the examination number. Each entry of the matrix indicates the number of students conflicting between the two examinations. Entry value of zero indicates no conflict between two exams, whereas a positive number means the existence of at least one conflict. This matrix facilitates managing different hard and soft constraints associated with timetabling and tracking the number of student enrolments in any pair of examinations.

B. Scheduling Module

This module is the central part of the overall system, aiming to generate a complete exam timetabling. Previous module passes necessary inputs so that scheduling module can produce appropriate exam timetabling based on user requirements. The scheduling of exams involves two steps. The first step is an initial feasible solution generation using an SD graph heuristic algorithm. In the second step, the quality of the solution is improved using local search algorithms. These include LAHC, SA, and GDA algorithms. Although not guaranteeing optimal solutions, they usually able to produce near-optimal solutions. Users can choose either of the steps for generating timetables and select desire algorithms with the appropriate parameters before execution. Note that improvement step does not execute independently, and it requires the solution vector of the first step. For example, improvement with the GDA algorithm phase is activated to produce a timetable when parameters such as decay rate and the number of iterations are properly defined, and initial feasible solutions produced by the graph heuristic algorithm are provided.

C. Reporting Module

Reporting module determines whether scheduling module is successfully generating the timetable or not according to the configuration assigned by users. Users can monitor the progress of the scheduling process, current penalty cost, execution time, and warning messages (if any) when scheduling module starts executing a timetable procedure. Once scheduling module produces a timetable, reporting module represents the final solution in a tabular form for the comprehension of everyone. Besides, subsidiary information such as total penalty costs (i.e., solution quality), execution times are also presented. At the end of the scheduling process, the document generator can be evoked to store final exam solutions (as an excel spreadsheet or pdf file format) in a disk for further uses.

D. User Interface

There is a graphical user interface that works at the top level. Users (e.g., students, teachers, human scheduler) can

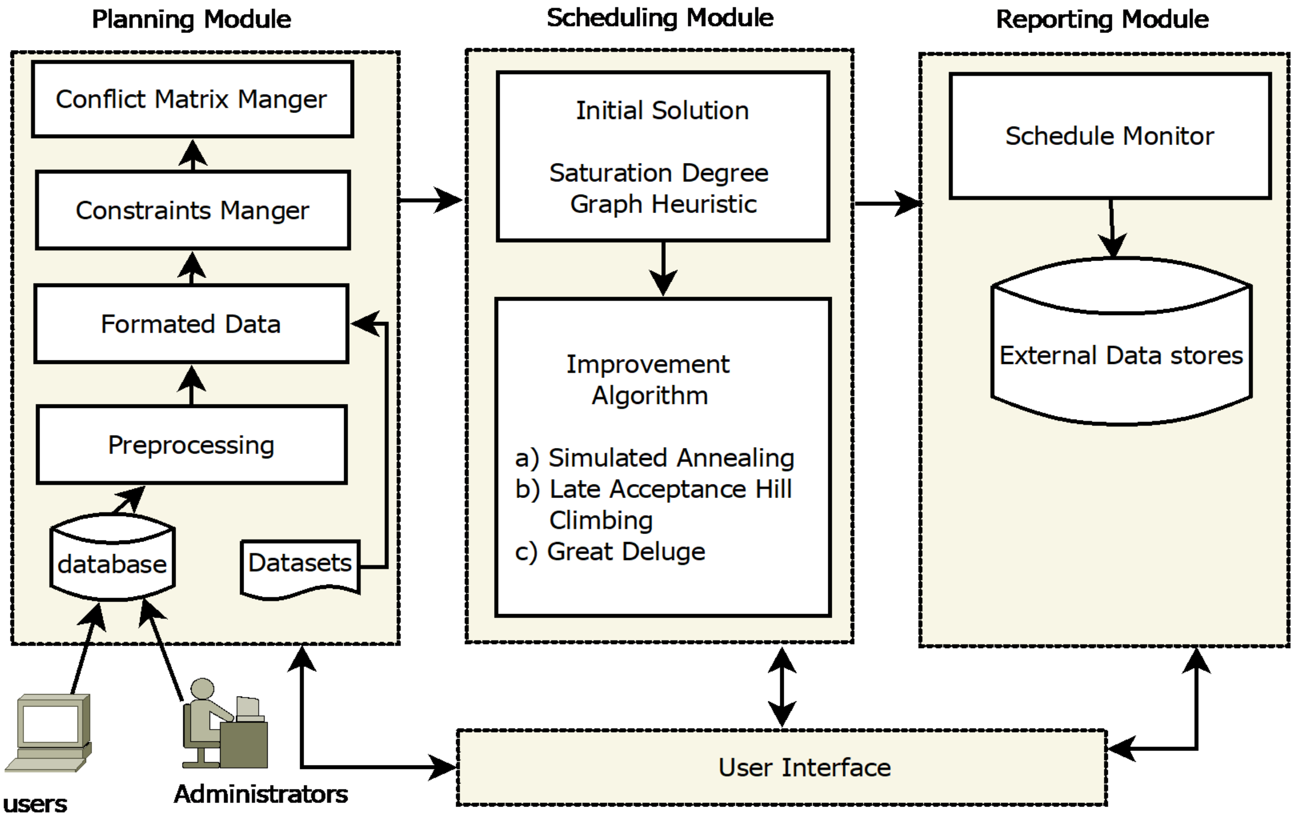


Fig. 1: Overview of system architecture

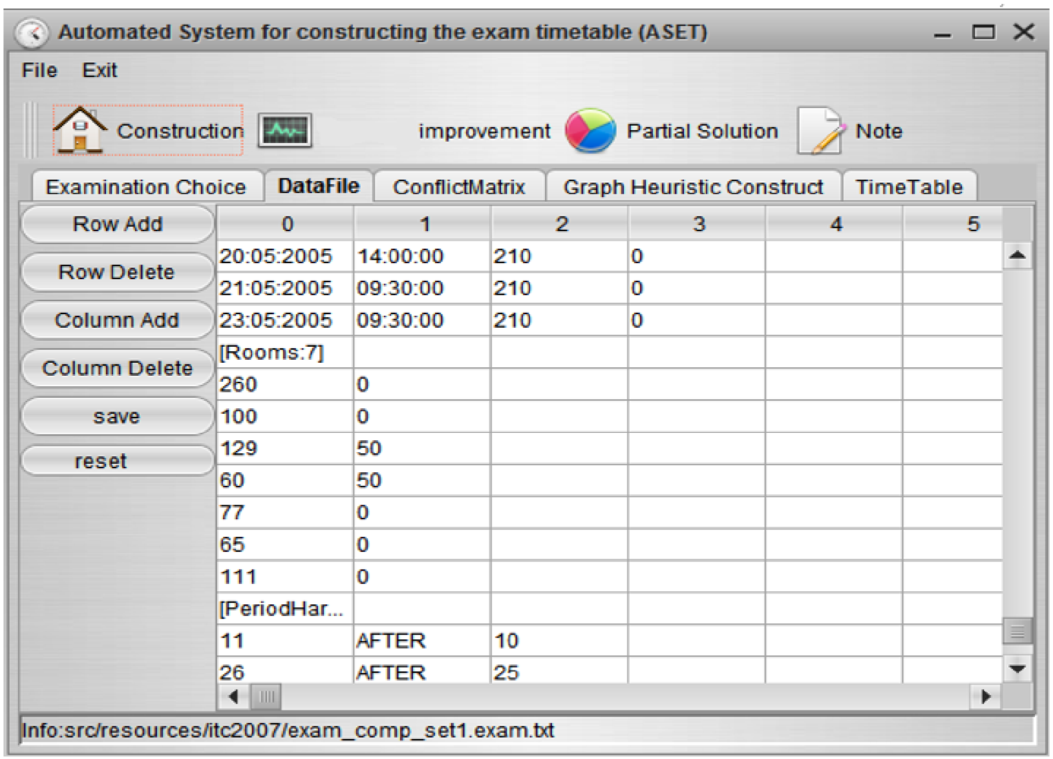
interact with all the modules using this interface. For example, visualizing the data, controlling the settings, and flow of the process can be done efficiently using this interface. Note that the whole system is developed using Java SE 1.7, and its swing library is deployed for GUI implementation.

1) Description of the system interface: The interface is designed in a straightforward way so that a user can operate the system with minimal effort and get better user experience. In addition to the graphical user interface, a command-line interface is provided for advanced users. As it is hard to illustrate all the options of the user interface, some selected screenshot of the GUI of automated exam timetabling is presented in Fig. 2, Fig. 3, and Fig. 4. The system window contains different key components, including a menu bar, toolbar, and tab panels. The menu bar at the top is used to load the dataset. Below the menu bar is a toolbar which provides important options such as construction and improvement phase of timetable. Different tab panels are associated with each tool, with each tab being used for performing different actions. For instance, the Data file tab of the construction tool contains data sets used for scheduling. Here users can customize a dataset, impose different constraints by six command buttons, and update other relevant information (see Fig. 2). A sample execution of an exam timetable process is shown in Fig. 3. There are two different types of components on the GUI that are used to monitor the status of the execution process. The text area shows the successful allocation of exams into the time slot and rooms (quality of the solution), and the progress bar indicates the percentage of progression of an exam timetable. The left side of the window contains some input fields (drop-down combo boxes, text boxes, etc.), which are used to select a search algorithm, tune parameters, and set stopping criterion. For example, for a sample execution with a GDA algorithm, users have to set the decay rate and the number of iterations. One may change the configuration and even run multiple executions simultaneously, as every execution of the scheduling process is an independent thread. Fig. 4 shows the presentation of the final examination timetable results in the data grid. Some performance measures, such as penalty cost (i.e., quality of timetabling) and the execution time are also displayed on the left side of data grid, and users can save and retrieve exam schedules for further uses.

V. Results and Discussion

Whether the timetabling system is viable in solving examination timetabling correctly, it has been tested with eight instances of ITC2007 exam dataset. In the experiment, graph heuristic SD is used for producing the initial solution, and three local search meta-heuristics are used individually for optimizing the initial solutions in the improvement phase. Three variations of the neighbourhood structures have been employed within an improvement algorithm. Their explanation is outlined as follows:

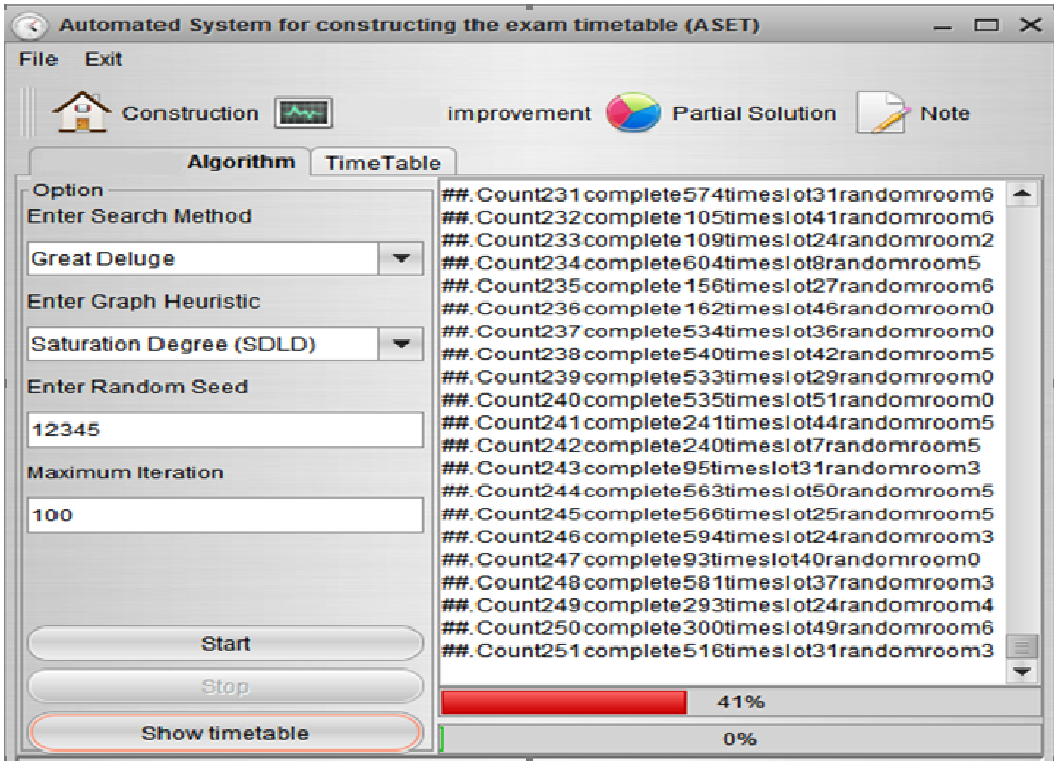
- $N1$: An examination is selected randomly and moves it to a random time slot.
- $N2$: Two examinations are selected randomly and swapping is occurred between their time slots.



0	1	2	3	4	5
20:05:2005	14:00:00	210	0		
21:05:2005	09:30:00	210	0		
23:05:2005	09:30:00	210	0		
[Rooms:7]					
260	0				
100	0				
129	50				
60	50				
77	0				
65	0				
111	0				
[PeriodHar...					
11	AFTER	10			
26	AFTER	25			

Info:src/resources/itc2007/exam_comp_set1.exam.txt

Fig. 2: Input with various entities



Enter Search Method: Great Deluge

Enter Graph Heuristic: Saturation Degree (SDLD)

Enter Random Seed: 12345

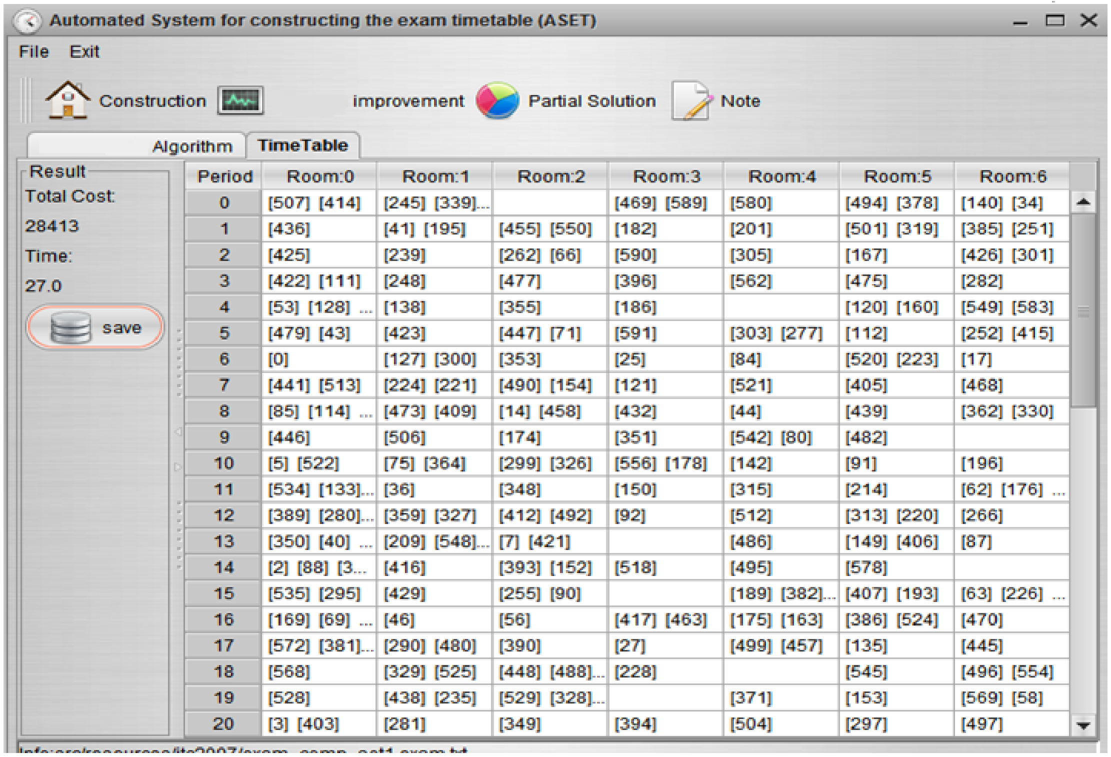
Maximum Iteration: 100

```
##.Count231complete574timeslot31randomroom6
##.Count232complete105timeslot41randomroom6
##.Count233complete109timeslot24randomroom2
##.Count234complete604timeslot8randomroom5
##.Count235complete156timeslot27randomroom6
##.Count236complete162timeslot46randomroom0
##.Count237complete534timeslot36randomroom0
##.Count238complete540timeslot42randomroom5
##.Count239complete533timeslot29randomroom0
##.Count240complete535timeslot51randomroom0
##.Count241complete241timeslot44randomroom5
##.Count242complete240timeslot7randomroom5
##.Count243complete95timeslot31randomroom3
##.Count244complete563timeslot50randomroom5
##.Count245complete566timeslot25randomroom5
##.Count246complete594timeslot24randomroom3
##.Count247complete93timeslot40randomroom0
##.Count248complete581timeslot37randomroom3
##.Count249complete293timeslot24randomroom4
##.Count250complete300timeslot49randomroom6
##.Count251complete516timeslot31randomroom3
```

41%

0%

Fig. 3: Executing of a timetabling with given parameters



Result
Total Cost:
28413
Time:
27.0

Period	Room:0	Room:1	Room:2	Room:3	Room:4	Room:5	Room:6
0	[507] [414]	[245] [339]...		[469] [589]	[580]	[494] [378]	[140] [34]
1	[436]	[41] [195]	[455] [550]	[182]	[201]	[501] [319]	[385] [251]
2	[425]	[239]	[262] [66]	[590]	[305]	[167]	[426] [301]
3	[422] [111]	[248]	[477]	[396]	[562]	[475]	[282]
4	[53] [128] ...	[138]	[355]	[186]		[120] [160]	[549] [583]
5	[479] [43]	[423]	[447] [71]	[591]	[303] [277]	[112]	[252] [415]
6	[0]	[127] [300]	[353]	[25]	[84]	[520] [223]	[17]
7	[441] [513]	[224] [221]	[490] [154]	[121]	[521]	[405]	[468]
8	[85] [114] ...	[473] [409]	[14] [458]	[432]	[44]	[439]	[362] [330]
9	[446]	[506]	[174]	[351]	[542] [80]	[482]	
10	[5] [522]	[75] [364]	[299] [326]	[556] [178]	[142]	[91]	[196]
11	[534] [133]...	[36]	[348]	[150]	[315]	[214]	[62] [176] ...
12	[389] [280]...	[359] [327]	[412] [492]	[92]	[512]	[313] [220]	[266]
13	[350] [40] ...	[209] [548]...	[7] [421]		[486]	[149] [406]	[87]
14	[2] [88] [3...	[416]	[393] [152]	[518]	[495]	[578]	
15	[535] [295]	[429]	[255] [90]		[189] [382]...	[407] [193]	[63] [226] ...
16	[169] [69] ...	[46]	[56]	[417] [463]	[175] [163]	[386] [524]	[470]
17	[572] [381]...	[290] [480]	[390]	[27]	[499] [457]	[135]	[445]
18	[568]	[329] [525]	[448] [488]...	[228]		[545]	[496] [554]
19	[528]	[438] [235]	[529] [328]...		[371]	[153]	[569] [58]
20	[3] [403]	[281]	[349]	[394]	[504]	[297]	[497]

Fig. 4: Desired exam timetable

- $N3$: Two time slots are selected randomly and all examinations between the two time slots are swapped.

The termination criteria for the improvement phase are fixed at 10000 iterations. Besides, 30 individual run is also performed for each instance. The following parameters are set for the local searches that are shown in Table III. Note that other setting of parameters could have been selected, but these parameter values have been selected according to the values used in the scientific literature.

TABLE III: Parameter Setting for experiments

Algorithm	Parameter	Value(s)
SA	Cooling rate	0.1
	Temperature	5000
LAHC	Frame size	500
GDA	Decay rate	0.1

From the experimental results in Table IV, it can be deduced that the proposed system can solve the ITC2007 exam benchmark dataset effectively. For all of the instances of the dataset, the best and the average values are highlighted after 30 individual runs. Graph heuristic SD produces feasible solutions for all of the instances, and local search approaches further improve the quality of the solutions. Among these three local search algorithms, 5 out of 8 cases (e.g., Exam_1, Exam_3, Exam_5, Exam_6, and Exam_7) LAHC performed the best results followed by GDA with instances Exam_2 and Exam_8, and SA with Exam_4. It is apparent that graph heuristic does not produce quality results because it considers only hard constraints. On the other hand, local search optimization algorithms produce better results compared to graph heuristics because the local search can improve the solution by reducing the soft constraint violations. Note that this paper does not aim to find the performance of the algorithms that suite the best. Instead, it shows the capabilities of graph heuristic and some local search algorithms for solving the exam timetable interactively. It is up to users to decide what will be the most useful algorithm in their particular circumstances. The advantages of the proposed interactive system are highlighted below:

- As the system has been developed using Java, it can run on a computer running both Windows and Linux (i.e., platform independence). Moreover, the incorporation of multi-thread assists the user to execute and analyze more than one exam instance simultaneously.
- This interactive tool is able to construct a complete timetable within a short time compared to the human scheduler, which usually takes long preparation in advance. It is also notable that maintaining hard constraints strictly and minimization of soft constraint violations for a large number of exams are challenging for the human scheduler. In contrast, an automated scheduler can perform the tasks firmly and effectively.
- This tool is useful as it can efficiently utilize institutional resources (i.e., room and timeslot utilization) as well as fulfill the major requirements requested by the

TABLE IV: Performance (penalty values) comparison between SD, LAHC, SA and GDA on ITC2007 exam dataset

Instances	**SD**		**LAHC**		**SA**		**GDA**	
Measure	**Best**	**Avg**	**Best**	**Avg**	**Best**	**Avg**	**Best**	**Avg**
Exam_1	25,989	26,769.45	12,421	13,048.88	12,537	14,042.37	12,483	13,858.68
Exam_2	30,960	32,135.67	2,807	3,766.79	2,911	3,553.26	2,789	3,578.79
Exam_3	85,356	88,374.43	43,098	47,160.20	44,173	48,060.62	43,241	47,560.63
Exam_4	41,702	42,323.38	34,241	34,937.07	34,152	34,834.18	34,417	34,744.46
Exam_5	132,953	133,873.50	15,643	16,773.46	15,816	16,891.51	15,690	16,612.17
Exam_6	44,160	48,729.67	29,630	33,880.08	30,116	34,910.16	29,845	34,150.38
Exam_7	53,405	56,366.08	19,080	21,612.42	20,071	21,518.31	19,178	21,821.75
Exam_8	92,767	96,465.32	23,315	25,002.29	23,411	25,801.79	22,891	25,152.65

students and invigilators. Institutional personnel can easily use and maintain the software without having prior programming knowledge.

- Flexibility in changing of input settings (e.g., constraints, exams, resources), support of interactive parameters tuning, and selection of different execution methods (either initial feasible solution with graph heuristic or initial solution followed by a near-optimum solution using local searches) are some salient features of the system, which makes it a robust interactive tool. Users can observe the effects of the different configurations on the output quality of the timetable.

- Although the system includes predefined eight instances of ITC2007 dataset, a provision has been kept for the users to modify or add new user-defined exam instances.

VI. Conclusions

This paper aims to generate an easy to use interactive examination timetabling software whereby graph heuristics and different local search algorithms are employed as solution methods. SD graph heuristic generates an initial feasible solution, whereas local search algorithms such as SA, GDA, and LAHC work as an optimizer for producing near-optimum solutions for the exam dataset. This proposed system is developed using Java and tested successfully on real-world dataset named ITC2007 exam dataset. The software is flexible and robust that outweighs the manual approaches. Users can automatically produce dataset from student registrations, select preferred hard and soft constraints, employ different improvement algorithms with desirable parameters and eventually produce a quality timetable within a reasonable time frame. The system could be scaled up by including different population-based search algorithms, which could provide more efficiency in the improvement phase. The usability of the software can also be enhanced to attain a satisfactory user experience.

Acknowledgment

This research is supported by Ministry of Posts, Telecommunication and Information Technology, Government of People's Republic of Bangladesh (Memo no: 56.00.0000.28.33.042.15-509)

References

[1] J. Johnes, "Operational research in education," *European Journal of Operational Research*, vol. 243, no. 3, pp. 683 – 696, 2015.

[2] P. Boizumault, Y. Delon, and L. Peridy, "Constraint logic programming for examination timetabling," *The Journal of Logic Programming*, vol. 26, no. 2, pp. 217 – 233, 1996.

[3] A. Cataldo, J.-C. Ferrer, J. Miranda, P. A. Rey, and A. Sauré, "An integer programming approach to curriculum-based examination timetabling," *Annals of Operations Research*, vol. 258, no. 2, pp. 369–393, 2017.

[4] N. R. Sabar, M. Ayob, R. Qu, and G. Kendall, "A graph coloring constructive hyper-heuristic for examination timetabling problems," *Applied Intelligence*, vol. 37, no. 1, pp. 1–11, 2012.

[5] M. Mohmad Kahar and G. Kendall, "A great deluge algorithm for a real-world examination timetabling problem," *Journal of the Operational Research Society*, vol. 66, no. 1, pp. 116–133, 2015.

[6] Y. Bykov and S. Petrovic, "A step counting hill climbing algorithm applied to university examination timetabling," *Journal of Scheduling*, vol. 19, no. 4, pp. 479–492, 2016.

[7] P. Amaral and T. C. Pais, "Compromise ratio with weighting functions in a tabu search multi-criteria approach to examination timetabling," *Computers & Operations Research*, vol. 72, pp. 160–174, 2016.

[8] M. Battistutta, A. Schaerf, and T. Urli, "Feature-based tuning of single-stage simulated annealing for examination timetabling," *Annals of Operations Research*, vol. 252, no. 2, pp. 239–254, 2017.

[9] N. Pillay and W. Banzhaf, "An informed genetic algorithm for the examination timetabling problem," *Applied Soft Computing*, vol. 10, no. 2, pp. 457–467, 2010.

[10] O. Abayomi-Alli, A. Abayomi-Alli, S. Misra, R. Damasevicius, and R. Maskeliunas, "Automatic examination timetable scheduling using particle swarm optimization and local search algorithm," in *Data, Engineering and Applications*, pp. 119–130, Springer, 2019.

[11] A. L. Bolaji, A. T. Khader, M. A. Al-Betar, and M. A. Awadallah, "A hybrid nature-inspired artificial bee colony algorithm for uncapacitated examination timetabling problems," *Journal of Intelligent Systems*, vol. 24, no. 1, pp. 37–54, 2015.

[12] H. Babaei, J. Karimpour, and A. Hadidi, "A survey of approaches for university course timetabling problem," *Computers & Industrial Engineering*, vol. 86, pp. 43–59, 2015.

[13] R. Qu, E. K. Burke, B. McCollum, L. T. Merlot, and S. Y. Lee, "A survey of search methodologies and automated system development for examination timetabling," *Journal of scheduling*, vol. 12, no. 1, pp. 55–89, 2009.

[14] S. Piechowiak and C. Kolski, "Towards a generic object oriented decision support system for university timetabling: an interactive approach," *International Journal of Information Technology & Decision Making*, vol. 3, no. 01, pp. 179–208, 2004.

[15] J. J. Thomas, A. T. Khader, B. Belaton, and E. Christy, "Visual interface tools to solve real-world examination timetabling problem," in *2010 Seventh International Conference on Computer Graphics, Imaging and Visualization*, pp. 167–172, IEEE, 2010.

[16] M. Ayob, A. R. Hamdan, S. Abdullah, Z. Othman, M. Z. A. Nazri, K. A. Razak, R. Tan, N. Baharom, H. A. Ghafar, R. M. Dali, *et al.*, "Intelligent examination timetabling software," *Procedia-Social and Behavioral Sciences*, vol. 18, pp. 600–608, 2011.

[17] Z. Chunbao and T. Nu, "An intelligent, interactive & efficient exam scheduling system (iieess v1. 0)," *Proceeding of the Practice and Theory of Automated Timetabling (PATAT), Norway*, pp. 437–450, 2012.

[18] I. Ober, "A variant of the high-school timetabling problem and a software solution for it based on integer linear programming," 2016.

[19] T. Müller, "Itc2007 solver description: a hybrid approach," *Annals of Operations Research*, vol. 172, no. 1, p. 429, 2009.

[20] B. McCollum, P. McMullan, A. J. Parkes, E. K. Burke, and R. Qu, "A new model for automated examination timetabling," *Annals of Operations Research*, vol. 194, no. 1, pp. 291–315, 2012.

[21] A. K. Mandal and M. Kahar, "Solving examination timetabling problem using partial exam assignment with great deluge algorithm," in *2015 International Conference on Computer, Communications, and Control Technology (I4CT)*, pp. 530–534, IEEE, 2015.

[22] E. K. Burke and Y. Bykov, "A late acceptance strategy in hill-climbing for exam timetabling problems," in *PATAT 2008 Conference, Montreal, Canada*, pp. 1–7, 2008.

[23] K. Bouleimen and H. Lecocq, "A new efficient simulated annealing algorithm for the resource-constrained project scheduling problem and its multiple mode version," *European journal of operational research*, vol. 149, no. 2, pp. 268–281, 2003.

[24] G. Dueck, "New optimization heuristics: The great deluge algorithm and the record-to-record travel," *Journal of Computational physics*, vol. 104, no. 1, pp. 86–92, 1993.

Invariant Feature Extraction for Component-based Facial Recognition

Adam Hassan[1]
Sudan University of Science and Technology
College of Computer Science & Information Technology
Khartoum, Sudan

Serestina Viriri[2]
University of KwaZulu-Natal
School of Maths, Statistics & Computer Science
Durban, South Africa

***Abstract*—This paper proposes an alternative invariant feature extraction technique for facial recognition using facial components. *Can facial recognition over age progression be improved by analyzing individual facial components?* The individual facial components: eyes, mouth, nose, are extracted using face landmark points. The Histogram of Gradient (HOG) and Local Binary Pattern (LBP) features are extracted from the individually detected facial components, followed by random subspace principal component analysis and cosine distance. One of the preprocessing steps implemented is the facial image alignment using angle of inclination. The experimental results show that facial recognition over age progression can be improved by analyzing individual facial components. The entire facial image can change over time, but appearance of some individual facial components is invariant.**

***Keywords*—Invariant features; facial components; facial recognition; age progression; HOG; LBP**

I. Introduction and Background

Face recognition is a challenging and relevant research area in image processing and computer vision community. Significant advances have been achieved throughout the last years. The majority of the research works studied general face recognition without considering face recognition over age progression. There are few research works which focus on age-invariant face recognition, and some of these related works are found in [21] [25] Hassan et al. [7] categorized the invariant discriminative feature extraction methods into two groups; generative models, and non-generative approaches. The generative models focus on learning the joint probability distribution while the non-generative models trade on the conditional probability distribution [17].

For age detection, an aging function is used and conducted with parametric model to get an exact age then another set of parameters are utilized to produce the target age [13]. However, aging transformations differ for different persons. An aging pattern subspace which deals with a sequence of individual face images, arranges them in time order that allows extracting features from both the shape and texture intensity [4]. Park et al. [20] used a model of 3D aging technique and show that the method can eliminate the age variations. However, generative methods used for face recognition suffer from poor aging process representation especially when just a few number of training image samples are available.

An effective alternative way to solve the limitations of generative models for better face representation is to implement local descriptors. Local descriptors are capable of recognizing aging variations, and are robust to additional intra-class variations recognition. Many techniques associated with local descriptors to discriminate aging variations features are proposed in [5] [11] [14] [23].

Among effective local feature descriptors, Local Binary Patterns (LBP) [1], [2] [18] emerges as the famous candidate method. However, it is not always that the uniform binary patterns have higher frequency as it is known but, sometimes non-uniform binary patterns perform better [6]. Gong et al. [6] considered the effectiveness of LBP, and proposed a feature descriptor associated with maximization of the code entropy. Another effective face descriptor of age progression is used for face verification described in [15] [16]. They proposed a method which uses the gradient orientation pyramid (GOP), and found that face recognition algorithms degrade when age gap exceeds four years.

Many approaches based on local features have been proposed. The local descriptors add an additional detailed facial characteristic that is substantial to the recognition process. Ahonen et al. [2] introduced LBP texture descriptor for facial representation. It is considered as an efficient and simple texture descriptor which labels the pixels of an image by thresholding the neighborhood of each pixel and considers the result as a binary number. The resulting descriptor formed from the histogram of the labels. Then histogram distance is used as dissimilarity measure between the pair of facial images.

Another extension of LBP is Enhanced Local Binary Patterns (ELBP) [24] which performs face representation using threshold constant to threshold pixels into three values. Another prominent feature representation approach is Elastic Bunch Graph Matching (EBGM) [26]. It is an algorithm which localizes a set of features of certain facial points and extracts Gabor jets at those points. One of high distinctive facial features representations is Scale Invariant Feature Transform (SIFT) [12] which serves feature invariance from different views: translation, scaling, and rotation. Bay et al. [3] introduced the Speeded Up Robust Features (SURF), which is a detector and descriptor for patented local feature. This method takes the advantage of distinctiveness, and robustness, so that can be calculated and compared much faster.

Moreover, biometric technology is one of the efficient methods to identify humans depending on face components characteristics. It is mentioned in the literature [27] that face recognition is considered one of the systems which is used in the field of biometric technology. Heisele et al. [10] extracted frontal face features to obtain eyes and the mouth components, then computed the triangle area between them. Component-based face recognition methods build upon multiple models with a number of facial components that represent an image on training phase. Many approaches are proposed on global-based technique, but component-base approaches are not intensively researched [8]. Geometry-based feature techniques for face recognition require calculation with geometrical features extracted from facial image. Face image representation can be viewed using size and position of facial components such as eyes, nose, mouth, forehead and cheeks.

The rest of the paper is organized as follows: Section II describes the methods and techniques implemented, Section III discusses the results achieved,and Section IV presents the conclusion and the envisioned future work.

II. Methods and Techniques

The proposed model consists of two main stages: preprocessing and feature extraction. Preprocessing stage consists of land mark face detection, face alignment, components cropping. The proposed model for invariant feature extraction for component-based facial recognition is depicted in Fig. 1 The proposed model allows all facial components to be allocate initial same weights, but their matching scores are sorted to select the highest score among them.

A. Preprocessing

Landmarks Detection: The original size of each image is 200x 240 pixels. At the preprocessing step, we initially detect 66 landmark points and use the two outer landmarks of eyes component to set the face image horizontally using specific calculated angle for rotation. When the image is well aligned according to eyes then detect new landmarks for next step which assist to crop facial components accurately. Fig. 2 shows the series of preprocessing steps.

Face recognition under pose variations is to recognize faces images of different poses. Face recognition rates are very poor when one tries to match images of different poses for the same person using any well known recognition technique. Hemlata et al. [9] detected eyes region using features of face connected components but, instead of that we use detected landmark points of outer corners of the two eyes to align the image face horizontally. if the face image is not horizontally aligned it must be rotated clock wise or anti clock wise base on angle direction. Angle of inclination of an image when it is correctly calculated then the image can be aligned. As depicted in Fig. 3, the two black filled circles represent the eyes, where:

$L(L_x, L_y)$ is the coordinates of left eye and $R(R_x, R_y)$ refers to the coordinates of right eye whereas the intersection of vertical and horizontal line makes the point M with its coordinates $M(L_x, R_y)$.

Opposite = $\sqrt{(Rx - Lx)^2 + (Ry–Ry)^2}$

Adjacent = $\sqrt{(Lx–Lx)^2 + (Ly–Ry)^2}$

To align facial image depicted in Fig. 2, the angle (theta) can be calculated using the equation:

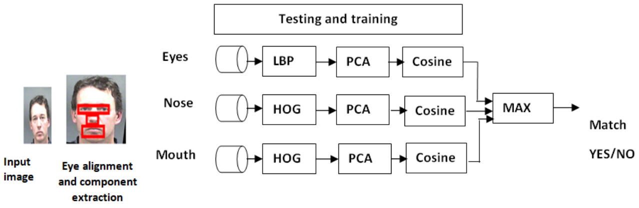


Fig. 1. Proposed Model for Component-based Facial Recognition

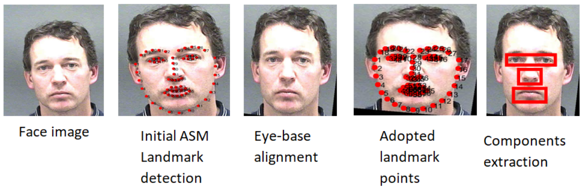


Fig. 2. Preprocessing steps

$$\theta = \tan\text{inverse}\,\frac{Opposite}{Adjacent} \tag{1}$$

Components Extraction: The work conducted on three facial components: eye pair, nose, and mouth. Eye pair component region extends from landmark point 37 to 43 as rectangle Length. To determine the width, arrange all the eye pair component points and select the highest and lowest points and the line between them is the rectangle width. For nose component, point 29 and the lowest nose point are selected besides point 32 to 36. The third component is the mouth where points 49 and 55 represent the length and highest and lowest points are the width.

B. Components Representation

This work implements face components matching using cosine distance for PCA features trained on two different separated descriptors. Rather than combining different features as in [19] we proposed to vote one descriptor feature per component as a result of pre-experiments which reduces computational cost and increases the overall performance. For all facial components, we extract both LBP and HOG features utilizing a patch size of $8x8$ without overlap. Features extracted from all patches are concatenated producing two different feature vectors for each facial component. However, we get different feature vector dimensions before passing it to PCA.

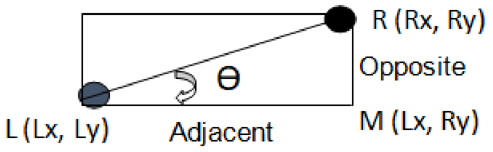


Fig. 3. Eye pair rectangle

LBP feature vector length is approximately four times the HOG feature vector length. Therefore, we sample each four LBP neighboring features as one feature point using mean value. The following equation shows how to form the desired features.

$$Desired\ features = \sum^{n} mean(Ri..j) \tag{2}$$

Where, n is the number of features is the raw features, i..j are the neighboring features. When the desired LBP features are obtained, then the feature dimensions are comparable. For each facial component, we build disjointed folds containing 300 genuine pairs each beside 40 different impostors. First, we start with age gap 0-1 to vote one descriptor that performs better than the other. Obviously, the age gap between each genuine pair is one year, as well as between the younger and impostor ones through all the forty impostors. We pass the features of one facial component to PCA for dimensionality reduction, and then we obtain PCA eigenvectors for genuine pairs and the corresponding impostor. Descriptor features of each component are treated separately till matching and decision phase then we fuse the scores of facial components.

C. Component Matching

For each pair one image (the younger one) is compared with all 40 impostor images- (each is 1 year older than the genuine one) producing 40 different measures. The decision is set to 1 If the cosine similarity between the two genuine pairs is less than all the 40 measures, otherwise, the decision is 0. As shown in Table I, we analyzed all age ranges and divide the data set into groups. while Table II shows the performance which is discussed in next section. Algorithm 1 describes the matching process. We construct fold from each group except the last two groups due to shortest of images numbers. All component decision scores are shared for final decision using the maximum operation as the following:

$$Final\ decision = \sum_{j=1}^{n} max(I_j) \tag{3}$$

Algorithm 1

```
for i= 1 to n do
    get the younger image of genuine pair get genuine(i)
    for j =1 to m do
        get impostor(j)
        measure(j) $\Leftarrow$ distance(genuine(i), impostor(j))
        measures $\Leftarrow$ [measures: measure(j)]
    end for
    sort all measures ascending measure Sort$\Leftarrow$ sort(measures)
    genuine Distance $\Leftarrow$ distance(genuine pair)
    if genuine Distance $<$ measure Sort
        match = 1 else match = 0
    end if
    matches $\Leftarrow$ [matches; match]
end for
```

The final decision is maximized if there exist only one component is observed as 1. However, this may lead to wrong decision if one component is false accepted while the two others are truly rejected, but this is not the case in this research work.

TABLE I. AGE RANGE GROUPS AND CORRESPONDING NUMBER OF IMAGES

Age range in years	No. of images
16–19	7469
20–29	16325
30–39	15354
40–49	12052
50–59	3599
60–69	319
70–77	16

TABLE II. PERFORMANCE PER EACH COMPONENT AND A NUMBER OF ALTERNATIVE COMBINATIONS

Age Gap	Eyes	Nose	Mouth	Eyes and Nose	Eyes and Mouth	Nose and Mouth	Eyes Nose and Mouth
0-1	81.21	83.89	80.54	96.64	92.61	95.97	97.98
1-5	77.41	80.67	70.85	94.04	88.61	90.57	95.65

III. RESULTS AND DISCUSSION

A. Data Set

We perform our work using the public domain MORPH [22] album 2 data set that contains 55,134 facial images from 13,617 classes. Table I shows the age ranges groups and corresponding number of images. We perform our work using two age gaps 0-1, and 1-5 and vote only one feature extractor for facial component. LBP and HOG feature extractors are extracted from each component then feature dimensions are intended to be most comparable, because we observed that the input feature dimension to PCA has crucial effect on accuracy. Equal dimensions are obtained as feature reduction for both extractors. Fig. 4 shows feature extractor performance for each component using ROC curves. Each component is analyzed using both extractors Therefore, we are entitled to choose the best for a later phase. Table II illustrates the best performance per each component and a number of alternatives established with combining two or three different components resulting in increasing the accuracy using final decision equation 2. Also we report combination of all components ROC carve for age gap 0-1 and gap 1-5 depicted in Fig. 5.

IV. CONCLUSIONS AND FUTURE WORK

We have analyzed different facial components over age progression utilizing a component based face representation and cosine similarity matching algorithm then max score for final decision. We perform our experiments on MORPH dataset and categorized it with different age ranges as in Table I. The proposed approach is robust to face recognition over

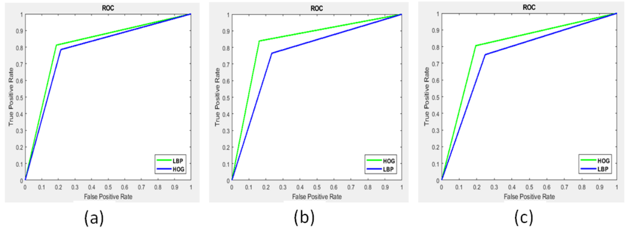


Fig. 4. Feature Extraction Voting per Component. (a) Eye pair: LPB performs better than HOG. (b) Nose: HOG perfoms better than LBP. (c) Mouth: else HOG is outperformed.

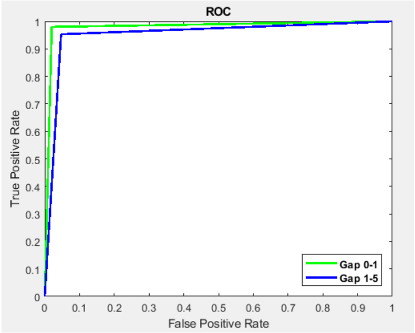


Fig. 5. MORPH results on 0-1 and 1-5 year age gap data sets

age progression. Our further work will include more facial components and additional dataset with larger age gaps to see how the proposed work can improve face recognition rate.

References

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face recognition with local binary patterns. In *European conference on computer vision*, pages 469–481. Springer, 2004.

[2] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):2037–2041, 2006.

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[4] Xin Geng, Zhi-Hua Zhou, Yu Zhang, Gang Li, and Honghua Dai. Learning from facial aging patterns for automatic age estimation. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 307–316. ACM, 2006.

[5] Dihong Gong, Zhifeng Li, Dahua Lin, Jianzhuang Liu, and Xiaoou Tang. Hidden factor analysis for age invariant face recognition. In *Proceedings of the ieee international conference on computer vision*, pages 2872–2879, 2013.

[6] Dihong Gong, Zhifeng Li, Dacheng Tao, Jianzhuang Liu, and Xuelong Li. A maximum entropy feature descriptor for age invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5289–5297, 2015.

[7] Adam Hassan and Serestina Viriri. Invariant feature extraction for facial recognition: A survey of the state-of-the-art. In *2018 Conference on Information Communications Technology and Society (ICTAS)*, pages 1–6. IEEE, 2018.

[8] Bernd Heisele, Thomas Serre, and Tomaso Poggio. A component-based framework for face detection and identification. *International Journal of Computer Vision*, 74(2):167–181, 2007.

[9] A Hemlata and Mahesh Motwani. Face detection by finding the facial features and the angle of inclination of tilted face. *International Journal of Computer Science Issues (IJCSI)*, 10(2 Part 1):472, 2013.

[10] Rein-Lien Hsu and Anil K Jain. Generating discriminating cartoon faces using interacting snakes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(11):1388–1398, 2003.

[11] Felix Juefei-Xu, Khoa Luu, Marios Savvides, Tien D Bui, and Ching Y Suen. Investigating age invariant face recognition based on periocular biometrics. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–7. IEEE, 2011.

[12] Dakshina Ranjan Kisku, Ajita Rattani, Enrico Grosso, and Massimo Tistarelli. Face identification by sift-based complete graph topology. In *2007 IEEE workshop on automatic identification advanced technologies*, pages 63–68. IEEE, 2007.

[13] Andreas Lanitis, Christopher J. Taylor, and Timothy F Cootes. Toward automatic simulation of aging effects on face images. *IEEE Transactions on pattern Analysis and machine Intelligence*, 24(4):442–455, 2002.

[14] Zhifeng Li, Unsang Park, and Anil K Jain. A discriminative model for age invariant face recognition. *IEEE transactions on information forensics and security*, 6(3):1028–1037, 2011.

[15] Haibin Ling, Stefano Soatto, Narayanan Ramanathan, and David W Jacobs. A study of face recognition as people age. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[16] Haibin Ling, Stefano Soatto, Narayanan Ramanathan, and David W Jacobs. Face verification across age progression using discriminative methods. *IEEE Transactions on Information Forensics and security*, 5(1):82–91, 2009.

[17] Andrew Y Ng and Michael I Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848, 2002.

[18] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):971–987, 2002.

[19] Charles Otto, Hu Han, and Anil Jain. How does aging affect facial components? In *European Conference on Computer Vision*, pages 189–198. Springer, 2012.

[20] Unsang Park, Yiying Tong, and Anil K Jain. Age-invariant face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):947–954, 2010.

[21] Narayanan Ramanathan, Rama Chellappa, and Soma Biswas. Computational methods for modeling facial aging: A survey. *Journal of Visual Languages & Computing*, 20(3):131–144, 2009.

[22] Tapan Kumar Sahoo and Haider Banka. Multi-feature-based facial age estimation using an incomplete facial aging database. *Arabian Journal for Science and Engineering*, 43(12):8057–8078, 2018.

[23] Diana Sungatullina, Jiwen Lu, Gang Wang, and Pierre Moulin. Multi-view discriminative learning for age-invariant face recognition. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013.

[24] Xiaoyang Tan and William Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE transactions on image processing*, 19(6):1635–1650, 2010.

[25] Yandong Wen, Zhifeng Li, and Yu Qiao. Latent factor guided convolutional neural networks for age-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4893–4901, 2016.

[26] Laurenz Wiskott, Jean-Marc Fellous, Norbert Krüger, and Christoph Von Der Malsburg. Face recognition by elastic bunch graph matching. In *International Conference on Computer Analysis of Images and Patterns*, pages 456–463. Springer, 1997.

[27] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458, 2003.

Feature Selection for Learning-to-Rank using Simulated Annealing

Mustafa Wasif Allvi[1], Mahamudul Hasan[2], Lazim Rayan[3], Mohammad Shahabuddin[4],
Md. Mosaddek Khan[5], Muhammad Ibrahim[6]
Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh[1,2,3,4]
Department of Computer Science and Engineering, University of Dhaka, Dhaka, Bangladesh[5,6]

Abstract—**Machine learning is being applied to almost all corners of our society today. The inherent power of large amount of empirical data coupled with smart statistical techniques makes it a perfect choice for almost all prediction tasks of human life. Information retrieval is a discipline that deals with fetching useful information from a large number of documents. Given that today millions, even billions, of digital documents are available, it is no surprise that machine learning can be tailored to this task. The task of learning-to-rank has thus emerged as a well-studied domain where the system retrieves the relevant documents from a document corpus with respect to a given query. To be successful in this retrieving task, machine learning models need a highly useful set of features. To this end, meta-heuristic optimization algorithms may be utilized. The aim of this work is to investigate the applicability of a notable meta-heuristic algorithm called simulated annealing to select an effective subset of features from the feature pool. To be precise, we apply simulated annealing algorithm on the well-known learning-to-rank datasets to methodically select the best subset of features. Our empirical results show that the proposed framework achieve gain in accuracy while using a smaller subset of features, thereby reducing training time and increasing effectiveness of learning-to-rank algorithms.**

Keywords*—*Information retrieval; learning-to-rank; feature selection; meta-heuristic optimization algorithm; simulated annealing

I. Introduction

Information retrieval (IR) is a process of retrieving the relevant information from a huge collection of data. Given the sheer amount of digital documents available today, this task is inherently quite difficult. An IR system works as follows. A query is submitted by the user of the system, and the task of the system is to return a ranked list of documents to the user based on the query. The user expects that highly relevant documents are in the top portion of the ranked list. Hence, the job of the system is to decide which document is relevant to the query and to what degree. To accomplish this task, researchers have been using heuristic scoring functions [1].

Machine learning can be thought of a discipline of applied statistics [2]. Given sufficient amount of empirical or historical data, these techniques are able to predict the outcome of unseen events. Various paradigms of machine learning are practised. Amongst these, supervised machine learning is mostly used by common people. In this setting, the training data consists of various information about different events along with the known labels. The job of the training module is to learn the pattern (in the form of a function) of the data that decides the labels. This function or model is then used to predict the labels of the unseen data, which is called the testing or evaluation module.

Learning-to-rank (LtR) is a relatively new area emerged in early 2000 as a successful marriage between information retrieval and machine learning [3]. In this framework, the training examples are query-document pairs, the features are the output scores of various scoring functions (such as tf-idf, bm25 score, etc.), and the labels are relevance scores assigned usually by humans. A model learnt from these data can then be used to generate relevance scores for documents with respect to a user's query. Fig. 1 depicts the scenario.

Today not only are data sets getting bigger and bigger, but also new data types have also been keeping to emerge, such as web-based data streams, genomics and proteomics micro arrays, and social media and system biology networks [4]. Therefore, the choice of features in a supervised machine learning setting is of utmost importance [5], [6]. One the one hand, we want to incorporate as much information as possible in our training set so that the learning algorithm can easily decide which aspects of the training data plays role in producing the labels. On the other hand, if irrelevant and misleading information is as features, the learning module will find it difficult to extract the pattern of the data, thereby reducing predictive accuracy. Moreover, if we can reduce the number of features, the training time in the learning phase will be minimized. For these reasons, a lot of research in supervised machine learning has been devoted to feature selection process [7]. The goals of feature selection include: creating easier and more comprehensible models, and enhancing data mining efficiency and helping to clean and understand data better [8]. It should be noted here that the feature selection is an NP-hard problem [9]. More details about these works will be elaborated in the next section.

The rest of the paper is organized as follows. In Section II, we briefly discuss existing works related to our area. In Section III, we discuss our framework in detail. Section IV presents the experimental settings and discusses the findings. Finally, Section V concludes the paper.

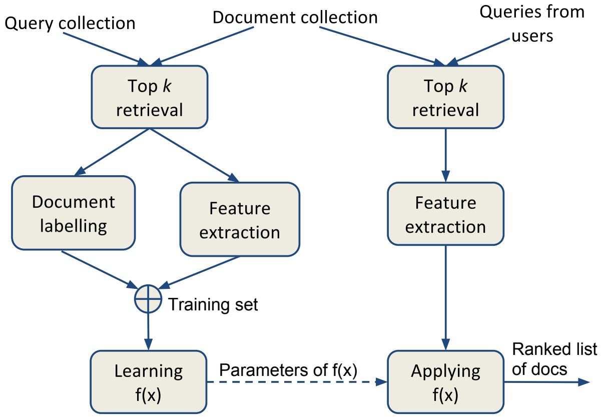


Fig. 1. An LtR-based IR system [10].

II. Background and Literature Review

In this section, we discuss the existing works related to our field which is feature selection in machine learning and learning-to-rank using traditional and meta-heuristic methods. We thus identify the gap in the existing literature.

A. Feature Selection in Supervised Machine Learning

We discuss the relevant papers of this subsection in two categories: (1) using traditional methods, and (2) using meta-heuristic algorithms.

1) Using Traditional Methods: By traditional methods of feature selection we mean filter, wrapper, embedded, forward/backward elimination, etc. methods [11].

Karegowda et al. [12] propose a supervised feature selection approach called wrapper approach. Wrappers take a subset of the function set, evaluates the output of the classifier on this subset, and then evaluates another subset on the classifier. Four different classifiers, namely Decision tree C4.5, Naïve Bayes, Bayes Network and Radial Basis are used. Eleven attributes identified by different wrappers were compared using different classifiers in the validation step. Their experiment discovers that no single standard wrapper approach is the best for different data sets.

Liu et al. [11] explain the importance of feature selection in data mining and briefly describe the methods of feature selection which is filter, wrapper and embedded model. For dimension reduction in data mining, the impact of feature selection is explained. There are brief explanation on feature weighting algorithms or subset selection, single data source algorithms, multi-source feature selection, detecting feature dependency, among other topics. Two research issues with selection features are explored.

Fan et al. [13] discusses the process of selecting features for data sets with millions of features.

2) Using Meta-Heuristic Algorithms: Heuristic optimization algorithms have been designed to solve large-scale optimization problems [14]. Most of these algorithms are nature-inspired. These methods differ from heuristic algorithms in the sense that the heuristic algorithms, working in a purely greedy manner, oftentimes stuck in local optima of the search space whereas the meta-heuristic algorithms use various types of randomization, even sometimes at the cost of apparently bad move, to get out of local optima. The focus of the meta-heuristic algorithms is to find an optimal solution for given problem by exploring maximum number of useful positions of the search space landscape in a given time frame.

Many researchers have investigated meta-heuristic algorithms for feature selection in machine learning. Below we describe some of them.

Emary et al. [15] propose a grey-wolf optimization (GWO) – a meta-heuristic algorithm inspired by natural instinct of wolves – based feature subset selection approach. Three parameters of the algorithm are used to decide the fitness of a candidate feature subset. The GWO algorithm iterates by exploring new regions within the function space and leverages solutions before near-optimal solution is reached. The optimization approach is based on k-nearest neighbor.

Sayed et al. [16] propose an extension of crow search algorithm. Features are selected based on the chaotic crow search algorithm (CCSA). CCSA is an upgraded version of crow search algorithm which is a nature based evolutionary algorithm. The paper use ten different chaotic maps for optimization. 20 data sets with different features and parameters are examined.

Aljarah et al. [17] explores the grasshopper optimization algorithm. A bio-inspired optimization technique is introduced to optimize the performance of Support Vector Machine (SVM) classifier, a powerful supervised machine learning technique. The model's main objective is to maximize SVM's classification accuracy with the minimum number of features. The suggested solution is tested on 18 public datasets and the result is satisfactory.

One of the most effective meta-heuristic algorithms to date is simulated annealing (SA) [18]. Being popular for its capability to find good quality solutions, SA is used by many researchers in multifarious machine learning domains. Gheyas and Smith [19] present a combination of two algorithms. The capability of better exploration in the search space of simulated annealing and the rapid convergence behavior of genetic algorithm are combined to find the feature subset more quickly and precisely. 11 synthetic and 19 real-world high-dimensional data sets to conduct the experiments.

Mafarja and Mirjalili [20] design a hybridization of whale optimization algorithm (WOA) and simulated annealing for feature selection. Whale optimization algorithm has some unique properties such as fewer parameters to control (since it requires only two key internal parameters to be modified), simple implementation and high versatility. The SA algorithm is wrapped with the WOA algorithm in an attempt to find the best solution throughout the neighborhood solutions. 18 data sets are examined for performance evaluation.

Barbu et al. [21] investigates feature selection methods using medical image data set where there is a massive number of features. The authors propose an algorithm that is suitable for big data computing due to its simplicity and ability to reduce the problem size throughout the iterations. The authors show that unlike its competitors such as boosting, the amount of data which the algorithm requires to use for training is much smaller, making it suitable for large-scale problems.

B. Feature Selection in Learning-to-Rank

Learning-to-rank has a wide area of application. Feature selection plays a vital role in building up a learning-to-rank model. Besides searching, image processing, big data learning and in classification of tweets are also included in the area of learning-to-rank implementation.

Novakovic et al. [22] combine ranking methods and classification algorithms to find the optimal feature subset. Four supervised algorithms, namely IB1, Naïve Bayes, C4.5 Decision Tree and Radia basis and statistical and entropy-based ranking methods are used. Filter-based methods are used for evaluating each subset of features, and irrelevant features are discarded. Duan et al. [23] use advanced greedy feature selection algorithm while exploring the earning-to-rank on tweets. Lai et al. [24] transform the feature selection problem into a joint convex optimization formulation which minimizes ranking errors as well as simultaneously conducting feature selection. Their framework can incorporate various feature similarity and imporance measures. Geng et al. [25] also pose the feature selection problem as an optimization problem by defining a loss function involving feature importance, and then solves it efficiently.

From the above review of existing works related the theme of this paper which is feature selection methods for learning-to-rank, we see that although simulated annealing have been studied to some extent for feature selection in supervised machine learning framework, to the best of our knowledge it has not been investigated in a learning-to-rank paradigm. Our work attempts to fill this gap in the literature.

III. Methodology

In this paper, we focus on finding the optimal subset of features of training data of LtR problem that is likely to yield a higher ranking accuracy during evaluation. For this optimization purpose, we utilize simulated annealing algorithm.

A. Simulated Annealing

Simulated annealing is usually used to solve NP-complete problems such as our feature subset selection problem, The algorithm basically combines two methods, namely hill climbing and random walk [14], [26].

Hill climbing is a greedy algorithm that only search for the local best solution. Hill climbing reaches a solution by recursively choosing the best neighbor based on an evaluation function, until there is no immediate better neighbor than the current one. When there is more than one best successor, a random selection is made from the set of best successors. For this nature of hill climbing algorithm, it often gets stuck in a local optimal point.

To overcome the problem of local optima of hill climbing algorithm, the idea of random walk is introduced [27]. The walk starts at a certain fixed node and moves randomly to a neighbor of the current node at each step. This method, however, has its own limitation because it may arbitrarily jump from one point to another in the search space.

Simulated annealing algorithm combines the merits of both hill climbing and random walk. It applies randomization in a way that allows occasional "bad" movements in an attempt to reduce the likelihood of getting stuck in a mediocre yet locally optimal solution. Specifically, the working procedure of simulated annealing is as follows. A random state is selected first. It then randomly selects a neighbor state (depending on the specific problem at hand, the definition of neighborhood is devised beforehand). If the selected neighbor state is better than the current state in terms of a utility function (again, decided beforehand), then the neighbor becomes the current state and the algorithm iterates over. But if the selected neighbor is worse than the current state, then the algorithm still gives it a chance to be selected as the (next) current state by a probability; the probability depends on the difference between the two states (current and neighbor) and the time during which the algorithm has been in its operation. More specifically, the higher the difference, the less the probability of choosing the (bad) neighbor, and the earlier stage the algorithm is in, the higher the said probability. Mathematically this probability is,

$$probability = e^{\frac{\Delta E}{T}},$$

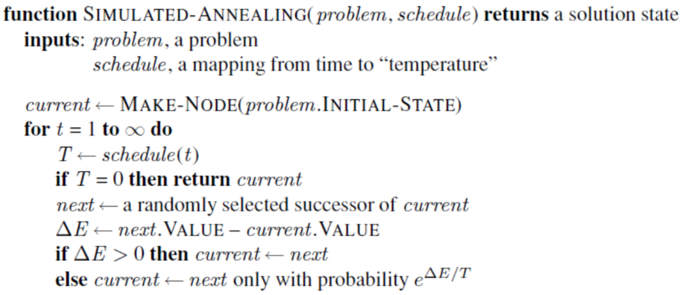
where,

$$\Delta E = neighbor\, state\, quality - current\, state\, quality,$$

and, T = temperature, which is reduced in every iteration from a very high value to zero. In essence, if the quality of neighbor state is worse than the current state, then the neighbor is selected (or not) with the probability $e^{\frac{\Delta E}{T}}$.

Although simulated annealing offers a way to overcome the striking bottleneck of hill climbing algorithm, but a large amount of time is oftentimes the price to be paid [28]. The idea of considering less value node is that, by doing it, the algorithm gets to explore more area in solution space by not getting stuck in a local optima.

Simulated annealing is known to work better than the bare local search algorithm most of the time. The procedure of simulated annealing is depicted in Fig. 2. To know more details about this exciting algorithm, the interested reader is requested to go though [29].



```
function SIMULATED-ANNEALING(problem, schedule) returns a solution state
  inputs: problem, a problem
          schedule, a mapping from time to "temperature"

  current ← MAKE-NODE(problem.INITIAL-STATE)
  for t = 1 to ∞ do
      T ← schedule(t)
      if T = 0 then return current
      next ← a randomly selected successor of current
      ΔE ← next.VALUE – current.VALUE
      if ΔE > 0 then current ← next
      else current ← next only with probability e^{ΔE/T}
```

Fig. 2. Simulated annealing algorithm [30]. "schedule" in line 3 is a monotonically decreasing function of T. "successor" in line 5, in our context, means neighbor. "Value" in line 6 implies quality of a state.

B. Proposed Framework

In this subsection we detail our algorithm that we use for selecting the best subset of features using simulated annealing technique. The procedure consists of broadly three constructs which are described below.

- *Notion of a state.* Here a state in the search space means a subset of features. Ultimately we search for the best subset of features that, when learnt using these features, yields the best predictive accuracy.
- *Definition of neighborhood.* A state's neighbor is defined as altering some features indexes of the current state randomly.[1]
- *Quality of a state expressed as a function.* Here we employ a heavily used IR evaluation metric called Normalized Discounted Cumulative Gain (NDCG) (will be elaborated in Section IV) as the quality of a state. The higher the NDCG of a model learnt from a training data (consisting of only the k features in question), the better.

Armed with these constructs, we are now in a position to detail our framework. Here is how it works. The procedure takes the number of features to select, k, as parameter, from the available pool of features. It then initially chooses random k features which is considered as the initial state in the search space. It then builds the training set using only these selected features, and trains an LtR algorithm on these data. The learnt model is then evaluated on test data (obviously using only the reduced subset of features in question) and stores the NDCG value as the quality of the current state. After that the neighbor state is chosen as per the rule aforementioned rule. The NDCG value in the evaluation stage is computed using the training data triggered by the neighbor state. The difference in these two NDCG values are then used to decide whether to make the neighbor state current state. The procedure stops when, for a particular k value, a predefined maximum number of iterations is reached. This entire procedure is repeated for various k values. While we retain and compare NDCG performance of various k values against corresponding random feature subset selection, ultimately the corresponding feature subset of highest performing k value that gives the best NDCG value across all the iterations is suggested to use instead of all available features. In our implementation we start with $k = 1$ and increase it one by one until we reach the number of available features.

IV. Empirical Results

For the experiments, we use six popular data sets, namely MQ2007, MQ2008, TD2004, HP2004, NP2004 and Ohsumed. These data sets have been made publicly available by Microsoft[2]. The data sets contain a varying number of features which are as follows: MQ2007 and MQ2008: 46, TD2004, HP2004, and NP2004: 64 and Ohsumed: 45. The data sets come with predefined chunks for training, validation and test sets. We maintain these chunks for the sake of better compatibility with the existing research works. To know more details about these datasets, please see [31], [32].

The RankLib LtR implementation[3] is used for evaluating our proposed framework. As for LtR algorithm, we choose LambdaMart because numerous research such as [33], [34] show that tree-ensemble methods in general, and specifically LambdaMart, perform oftentimes better than other LtR algorithms.

As mentioned earlier, as evaluation metric we use NDCG. NDCG is a ranking performance evaluation metric that gives a gradually higher score (out of 1) to a list having the highly relevant documents in the top portion of it. To know more about NDCG and other IR evaluation metric, please see [3].

For each of the six data sets, we generate a plot as follows. For a particular k value starting from 1 and ending in the number of available features, we generate two new training sets – one with the features suggested by the random selection process (i.e., by choosing random k feature indexes) and the other with the features prescribed by simulated annealing algorithm (i.e., by choosing k features of the solution state returned by the SA algorithm after 100 iterations). The LambdaMart LtR algorithm is then learnt for each of the generated training sets, and then the two learnt models are evaluated on the test set (obviously comprising with the same features of each case). Thus we get two NDCG values for a particular k value: one for the random selection process and the other for the SA algorithm. This way all possible k values are examined, and finally the graph is generated (for this data set) by plotting these two curves. Fig. 3, 4, 5, 6, 7, and 8 show these plots for the six data sets. Now, we analyze each of the six plots in details.

1) MQ2007 Data Set: Fig. 3 shows that the curve fluctuate initially at a higher degree which is gradually quelled. This is because when the number of features is small, the benefit of SA algorithm is not evident. Then as the number of features increases the curve get relatively flatter. From the graph we can say that instead of taking all the 46 features, we can get the best accuracy with only 26 features as prescribed by the SA algorithm

2) MQ2008 Data Set: Fig. 4 shows that for MQ2008 data set from the beginning the efficacy of SA approach is evident. With almost only 20 features suggested by the SA algorithm we can reach the accuracy of the training set having all 46 features.

3) Ohsumed Data Set: Fig. 5 shows that the curve of Ohsumed data set tends to vary almost from beginning to end for all the k values. It can be concluded that initially there is a significant difference between the random selection score and SA score. But after coming to about 18 features, the gap diminishes onward. Therefore, from our experiment we can say that using a subset of 18 features instead of all 45 features may be time-efficient.

4) NP2004 Data Set: Fig. 6 demonstrates that from around $k = 8$, i.e., using only 8 or more features with simulated annealing is almost invariably better than the random selection

[1] We note here that there could be other definitions which we intend to investigate as future work.

[2] https://www.microsoft.com/en-us/research/project/letor-learning-rank-information-retrieval/.

[3] https://sourceforge.net/p/lemur/wiki/RankLib/

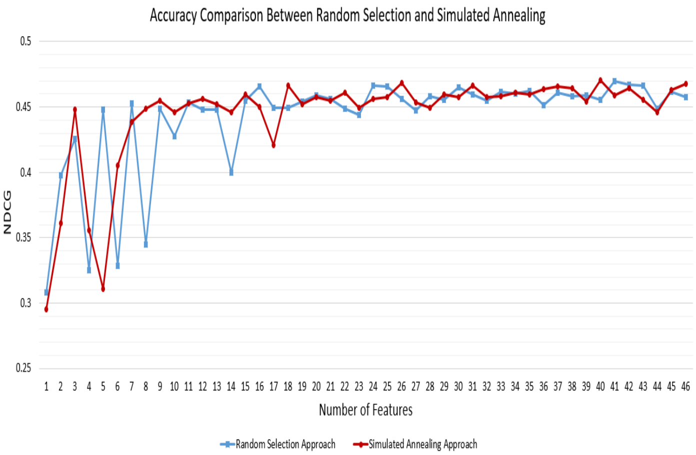


Fig. 3. MQ2007 data: comparison between random selection and simulated annealing in terms of NDCG.

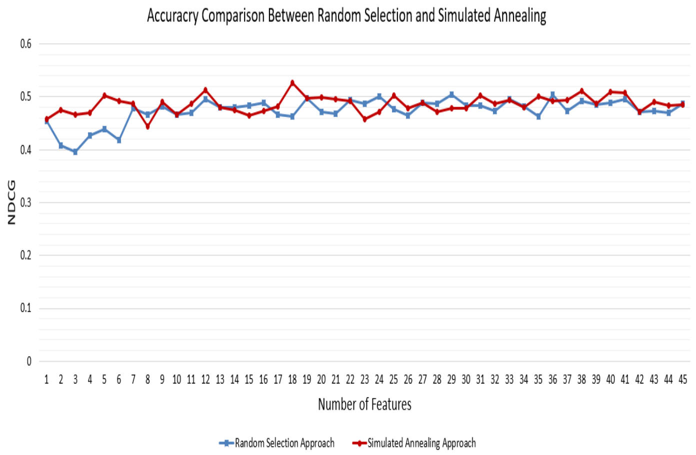


Fig. 5. Ohsumed data: Comparison between random selection and simulated annealing in terms of NDCG.

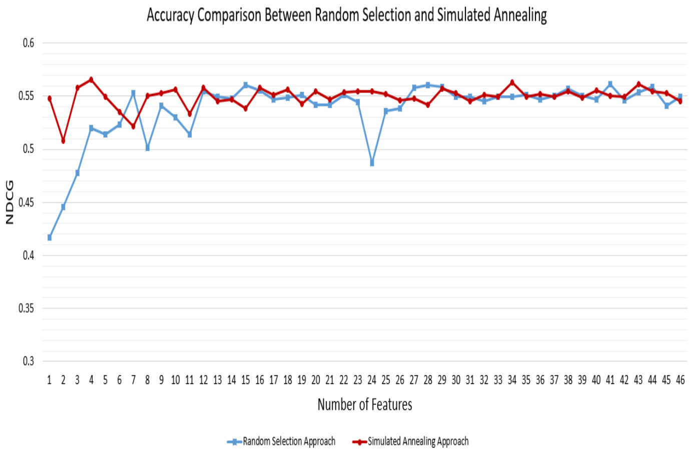


Fig. 4. MQ2008 data: comparison between random selection and simulated annealing in terms of NDCG.

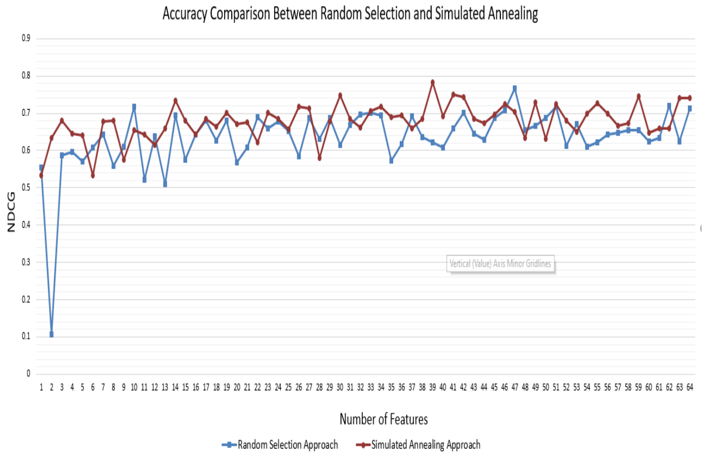


Fig. 6. NP2004 data: comparison between random selection and simulated annealing in terms of NDCG.

approach. Moreover, using 39 features outperforms the setting of using all 64 features.

5) HP2004 Data Set: Fig. 7 demonstrates similar trend that of NP2004 in terms of performance comparison between random selection and simulated annealing approaches. In particular, after $k = 8$ it appears that SA approach almost always outperforms the random selection approach. Taking a subset of only around 22 features is likely to beat the performance of 53 features.

6) TD2004 Data Set: Fig. 8 shows largely similar trend to that of HP2004 and NP2004. From around 8 features, the SA approach seems to outperform the random selection approach. Moreover, it appears that using only around 23 features yield equivalent accuracy to that of using all 64 features.

A. Discussion

The following points can be drawn from the analysis of experimental results.

- For all six data sets, a smaller subset of features work quite well as compared to the setting of using all available features. This indicates that blindly incorporating as many features as possible may not improve the accuracy of LtR systems, rather a careful selection of features is needed.
- Broadly, all six data sets appear to reap benefit of the simulated annealing feature selection method over the random selection method. It should be noted that we have used a basic SA algorithm. Recent variations of SA algorithm and other meta-heuristic algorithm may yield further improvement.
- Fluctuation is present in all the plots which is natural given the simulated annealing is a randomized algorithm. If we average the results of several runs, the fluctuation will be minimized.
- The nature of the features selected by the random selection and simulated annealing based selection has not been investigated.

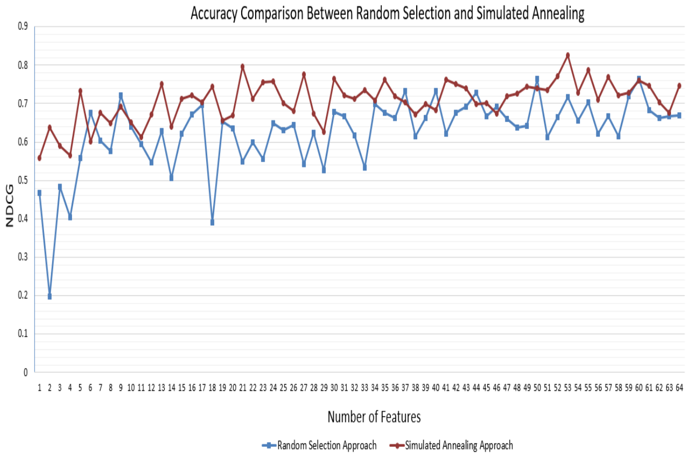


Fig. 7. HP2004 data: Comparison between random selection and simulated annealing in terms of NDCG.

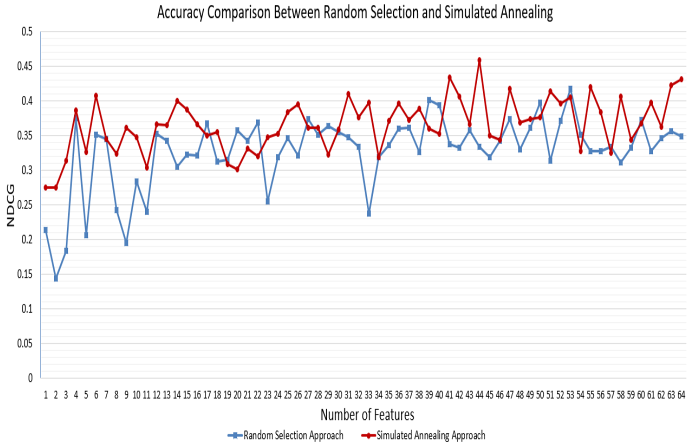


Fig. 8. TD2004 data: Comparison between random selection and simulated annealing in terms of NDCG.

- More iterations in simulated annealing may uncover further insights into our findings.

In essence, we can say that if our proposed framework is deployed, we are able to not only discover better feature subset to learn an LtR model but also to reduce the training time. This investigation thus suggests that commercial IR systems such as search engines that deploy LtR system may apply feature selection methods more seriously and wisely using sophisticated techniques like simulated annealing.

V. Conclusions and Future Work

Recently the area of ranking in information retrieval has earned a lot of attention in the field of machine learning. Learning-to-rank paradigm that is a blend between information retrieval and supervised machine learning has gained much momentum in the research community due to the success in satisfying users of information retrieval systems such as search engines. Performance of these algorithms heavily depend on the features or attributes used in the learning module. Hence a careful selection of features is needed. In this work we have deployed a effective and efficient meta-heuristic algorithm called simulated annealing to select a better subset of features from the available ones. Our experiments on benchmark data sets reveal that using simulated annealing we can extract an effective yet smaller subset of features that performs quite well as compared to the baseline (i.e., the setting where all features are used). This investigation suggests that the features should be chosen carefully so as to improve the predictive accuracy of the LtR algorithms as well as to reduce training time.

This work generated at least three-pronged future research avenues. Firstly, in this work we have examined only one, albeit highly effective and hence popular, LtR algorithm. It is natural to be curious about performance of other LtR algorithms when plugged in into our proposed framework. Secondly, larger LtR data sets need to be investigated. Thirdly, while we have explored the classical simulated annealing algorithm, other contemporary meta-heuristic algorithms may deserve such investigation.

References

[1] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge university press, 2008.

[2] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[3] M. Ibrahim and M. Murshed, "From tf-idf to learning-to-rank: An overview," in *Handbook of Research on Innovations in Information Retrieval, Analysis, and Management*. IGI Global, 2016, pp. 62–109.

[4] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu, "Advancing feature selection research," *ASU feature selection repository*, pp. 1–28, 2010.

[5] J. Brownlee, "An introduction to feature selection," *Machine Learning Process*, vol. 6, 2014.

[6] V. Kumar and S. Minz, "Feature selection: a literature review," *SmartCR*, vol. 4, no. 3, pp. 211–229, 2014.

[7] T. R. Brick, R. E. Koffer, D. Gerstorf, and N. Ram, "Feature selection methods for optimal design of studies for developmental inquiry," *The Journals of Gerontology: Series B*, vol. 73, no. 1, pp. 113–123, 2018.

[8] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu, "Advancing feature selection research," *ASU feature selection repository*, pp. 1–28, 2010.

[9] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Advances in bioinformatics*, vol. 2015, 2015.

[10] M. Ibrahim and M. Carman, "Comparing pointwise and listwise objective functions for random-forest-based learning-to-rank," *ACM Transactions on Information Systems (TOIS)*, vol. 34, no. 4, p. 20, 2016.

[11] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, "Feature selection: An ever evolving frontier in data mining," in *Feature selection in data mining*, 2010, pp. 4–13.

[12] A. G. Karegowda, M. Jayaram, and A. Manjunath, "Feature subset selection problem using wrapper approach in supervised learning," *International journal of Computer applications*, vol. 1, no. 7, pp. 13–17, 2010.

[13] J. Fan, R. Samworth, and Y. Wu, "Ultrahigh dimensional feature selection: beyond the linear model," *Journal of machine learning research*, vol. 10, no. Sep, pp. 2013–2038, 2009.

[14] J. Rajpurohit, T. K. Sharma, A. Abraham, and A. Vaishali, “Glossary of metaheuristic algorithms,” *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 9, pp. 181–205, 2017.

[15] E. Emary, H. M. Zawbaa, and A. E. Hassanien, “Binary grey wolf optimization approaches for feature selection,” *Neurocomputing*, vol. 172, pp. 371–381, 2016.

[16] G. I. Sayed, A. E. Hassanien, and A. T. Azar, “Feature selection via a novel chaotic crow search algorithm,” *Neural computing and applications*, vol. 31, no. 1, pp. 171–188, 2019.

[17] I. Aljarah, A.-Z. Ala’M, H. Faris, M. A. Hassonah, S. Mirjalili, and H. Saadeh, “Simultaneous feature selection and support vector machine optimization using the grasshopper optimization algorithm,” *Cognitive Computation*, vol. 10, no. 3, pp. 478–495, 2018.

[18] P. J. Van Laarhoven and E. H. Aarts, “Simulated annealing,” in *Simulated annealing: Theory and applications*. Springer, 1987, pp. 7–15.

[19] I. A. Gheyas and L. S. Smith, “Feature subset selection in large dimensionality domains,” *Pattern recognition*, vol. 43, no. 1, pp. 5–13, 2010.

[20] M. M. Mafarja and S. Mirjalili, “Hybrid whale optimization algorithm with simulated annealing for feature selection,” *Neurocomputing*, vol. 260, pp. 302–312, 2017.

[21] A. Barbu, Y. She, L. Ding, and G. Gramajo, “Feature selection with annealing for computer vision and big data learning,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 2, pp. 272–286, 2016.

[22] J. Novaković, “Toward optimal feature selection using ranking methods and classification algorithms,” *Yugoslav Journal of Operations Research*, vol. 21, no. 1, 2016.

[23] Y. Duan, L. Jiang, T. Qin, M. Zhou, and H.-Y. Shum, “An empirical study on learning to rank of tweets,” in *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 295–303.

[24] H.-J. Lai, Y. Pan, Y. Tang, and R. Yu, “Fsmrank: Feature selection algorithm for learning to rank,” *IEEE transactions on neural networks and learning systems*, vol. 24, no. 6, pp. 940–952, 2013.

[25] X. Geng, T.-Y. Liu, T. Qin, and H. Li, “Feature selection for ranking,” in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 407–414.

[26] D. S. Johnson, C. R. Aragon, L. A. McGeoch, and C. Schevon, “Optimization by simulated annealing: An experimental evaluation; part i, graph partitioning,” *Operations research*, vol. 37, no. 6, pp. 865–892, 1989.

[27] C. Avin and B. Krishnamachari, “The power of choice in random walks: An empirical study,” in *Proceedings of the 9th ACM international symposium on Modeling analysis and simulation of wireless and mobile systems*, 2006, pp. 219–228.

[28] E.-G. Talbi and T. Muntean, “Hill-climbing, simulated annealing and genetic algorithms: a comparative study and application to the mapping problem,” in *[1993] Proceedings of the Twenty-sixth Hawaii International Conference on System Sciences*, vol. 2. IEEE, 1993, pp. 565–573.

[29] D. Henderson, S. H. Jacobson, and A. W. Johnson, “The theory and practice of simulated annealing,” in *Handbook of metaheuristics*. Springer, 2003, pp. 287–319.

[30] S. Russell and P. Norvig, “Artificial intelligence: a modern approach,” 2002.

[31] M. Ibrahim, “Reducing correlation of random forest–based learning-to-rank algorithms using subsample size,” *Computational Intelligence*, vol. 35, no. 4, pp. 774–798, 2019.

[32] ——, “Sampling non-relevant documents of training sets for learning-to-rank algorithms,” *International Journal of Machine Learning and Computing*, vol. 10, no. 3, pp. 1–10, 2020 (In Press).

[33] C. J. Burges, “From ranknet to lambdarank to lambdamart: An overview,” *Learning*, vol. 11, no. 23-581, p. 81, 2010.

[34] M. Ibrahim, “An empirical comparison of random forest-based and other learning-to-rank algorithms,” *Pattern Analysis and Applications*, pp. 1–23, 2019.

Software-Defined Networking (SDN) based VANET Architecture: Mitigation of Traffic Congestion

Tesfanesh Adbeb[1]
School of Computer Science and Technology
Dalian University of Technology
Dalian, China

Wu Di[2]
School of Computer Science and Technology
Dalian University of Technology
Dalian, China

Muhammad Ibrar[3]
School of Software
Dalian University of Technology
Dalian, China

***Abstract*—In VANETs (Vehicular Ad-hoc Networks), the number of vehicles increased continuously, leading to significant traffic problems like traffic congestion, a feasible path, and associated events like accidents. Though, the Intelligent Transportation System (ITS) providing excellent services, such as safety applications and emergency warnings. However, ITS has limitations regarding traffic management tasks, scalability, and flexibility because of the enormous number of vehicles. Therefore, extending the traditional VANET architecture is indeed a must. Thus, in the recent period, the design of the SD-VANETs (Software-Defined Networking defined VANETs) has gained significant interest and made VANET more intelligent. The SD-VANET architecture can handle the aforesaid VANET challenges. The centralized (logically) SDN architecture is programmable and also has global information about the VANET architecture. Therefore, it can effortlessly handle scalability, traffic management, and traffic congestion issues. The traffic congestion problem leads to longer trip times, decreases the vehicles' speed, and prolong average end-to-end delay. Though, somewhere, some routes in the network are available with capacity, which can minimize the congestion problem and its characteristics. Therefore, we proposed heuristic algorithms called Congestion-Free Path (CFP) and Optimize CFP (OCFP), in SD-VANET architecture. The proposed algorithms address the traffic congestion issue and also provide a feasible path (less end-to-end delay) for a vehicle in VANET. We used the NS-3 simulator to evaluate the performance of the proposed algorithms, and for generating a real scenario of VANET traffic; we use the SUMO module. The results show that the proposed algorithms decrease road traffic congestion drastically compared to exiting approaches.**

Keywords*—*Software-Defined Networking; VANET; congestion; feasible path; NS3; SUMO

I. Introduction

In the current communication era, VANETs have received the significant attention of the researchers because of its unique and critical characteristics like frequent changes in topology, link failure, network stability, efficient traffic management, safety, congestion, and reliability [1], [2]. The characteristics, as mentioned earlier, lead to network instability because of high vehicle mobility. Thus, high vehicle mobility yields overall network efficiency, create road-side safety, and security issues. Therefore, ITS deployment is required to handle the enormous traffic efficiently, avoid congestion, reliability, and also provide the services to the passengers (like safety applications, emergency warnings, video streaming, lane change warning, and entertainment). These types of services, as mentioned before, need efficient and improved Packet Delivery Ratio (PDR), need high-quality communication, congestion-free path, and average end-to-end delay.

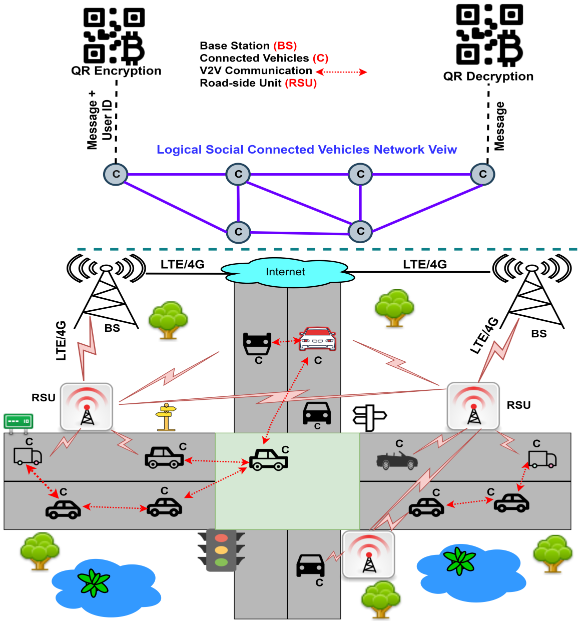


Fig. 1. V2X Architecture

ITS is a vital next-generation transportation system [3], [4], and it is a combination of communication technologies used in VANET management (i.e., efficiency, safety, and sustainability) and leading-edge information. In VANET, vehicles are like mobile nodes. They collect and disseminate information about their speed, current position, destination [5], [6], [7], [8]. In some emergency conditions, such are health issues, road accidents, and congestion, the VANET architecture (ITS) ensure the driving safety, alternative routes, and timely report. Therefore, through V2X (Vehicle-to-Everything) [9] architecture, it is possible to inform the nearby vehicles in a specific area to avoid congestion, emergency conditions, and provide the alternate reliable route, as shown in Fig. 1. Noticeably, under some traffic conditions, the shortest route can lead to the congestion problem. The congestion problem leads to longer trip times, decreases the vehicle's speed, and prolong end-to-end delay [10]. Although, somewhere, some routes in the network are available with capacity, which can minimize the congestion problem and its characteristics. The

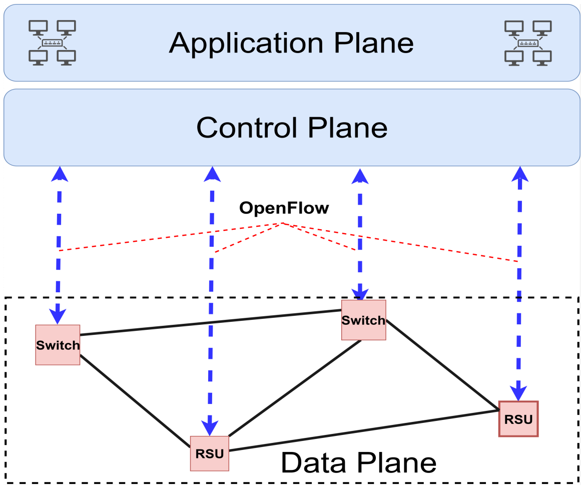


Fig. 2. SDN-based Architecture

goal of traffic engineering is to make sure that traffic is managed such that community ability is utilized efficiently and in a balanced manner. There are several techniques to handle Traffic Engineering (TE) problems like congestion and delay in VANETs.

To overcome the congestion problem in VANETs and its characteristics (i.e., prolong end-to-end delay, longer trip times, an emergency condition), we look at controlled (logically) SDN architecture. The programmable SDN architecture provides a flexible way to manage and control the traditional VANET architecture systematically. The main objective of programmable and logically centralized SDN architecture is to decouple the control plane from the data plane [11], [12], [13]. In SDN architecture, the unified controller (control plane) is responsible for monitoring, controlling, and managing the network resources efficiently. The purpose of the SDN controller is to improve and optimize the overall network performance like path selection, traffic control, congestion control, and efficient communication. The data plane is a networking infrastructure, forwarding devices (switch, router, Access Point (AP), and Roads-side Unit (RSU)) used for the data forwarding process. These forwarding devices connect with wired or wireless channels.

The SDN controller uses the OpenFlow protocol [11] for communication with data forwarding devices, as shown in Fig. 2. The proposed SD-VANET architecture consists of the following components. (a) Control Plane: the logically unified SDN controller that provides control functionalities about the entire network. The controller communicates with network devices in the data plane, and with the application. (b) Data Plane: The data plane consists of the vehicles (i.e., SDN-enabled wireless mobile nodes) and RSU (i.e., SDN-enabled stationary nodes) that receive the control message from the control plane. Each SDN-enabled wireless node contains a local agent, called SDN-agent. The SDN-agent is used to communicate with the controller.

The current approaches focused on the performance of SD-VANET routing, link stability, heterogeneity, offloading, and mobility. However, these approaches do not distribute the vehicles based on vehicles' density proactively; therefore, which causes congestion and prolongs the end-to-end delay. Congestion also leads to traffic accidents [1], [2], [13], [5] and also decreases the packet delivery ratio and QoS (Quality-of-Services). Motivated by the congestion problem and its characteristics in VANET, as mentioned earlier, we proposed new heuristic algorithms called Congestion Free Path (CFP) and Optimized CFP (OCFP) in the SD-VANET architecture. More explicitly, in this paper, our main contributions are summarized are follows.

- Prominently, the centralized (logically) SDN controller can manage, control, and provide flexible communication between V2I (Vehicle-to-Infrastructure) and V2V (Vehicle-to-Vehicle). More specifically, in SD-VANET architecture, from the global (abstract) view of the VANET, the centralized controller compute the congestion-free and optimal path for a vehicle.
- In this paper, we proposed heuristic algorithms called CFP and OCFP in the SD-VANET architecture. More specifically, these algorithms proactively distribute the vehicles on roads according to vehicles' density and compute the congestion-free path and its characteristics such as long queuing delay, longer trip times, decrease the vehicle's speed, and safety.
- The results of the proposed algorithms show that it decrease the traffic congestion ratio significantly.
- To evaluate the proposed algorithms, in this paper, we use NS-3.29 and SUMO. The SUMO is used to generate realistic road scenarios.

The remainder of the paper is systematized as follows. Section II categorized the related work into two subcategories, such as traditional schemes for road traffic congestion mitigation and SD-VANET architecture. The importance of the problem statement is shown in III. We present the proposed solution in Section IV to compute the feasible path (the congestion-free path with minimum end-to-end delay) for a vehicle in the SD-VANET architecture. Section V presents the experimental setup and simulation results of the proposed algorithms. In last, Section VI shows the conclusion.

II. Related Work

In this section, the related work is categorized into two categories, as follows.

A. Scheme for Congestion Mitigation and Detection

Recently one of the foremost research topics in VANETs is solving the problem of traffic congestion. The approaches proposed focused on V2V or V2I communications pose significant limitations below.

The authors proposed a scheme named TrafficView [10]. The TrafficView scheme considers the aggregated dissemination of traffic information such as broadcast time, average speed, and position of the specific vehicle on the road. This, in turn, a vehicle can learn about other on-road vehicles. However, this scheme does not consider the congestion and road traffic safety problem. In SOTIS (Self Organizing Traffic Information System) [14], the process of information exchange would be like [10], in which vehicles regularly relay data about themselves and other vehicles that they sense. In [15], the proposed scheme applies to the prediction and simulation of traffic congestion algorithms in distributed V2V architecture. In this scheme, if the travel time of a vehicle surpasses the consistent travel time in a free-flow condition, then the route is considered as a congested route. Additionally, for experimental purposes, the authors in the proposed work, monitor each road lane for a day. In the proposed scheme, the centralized entity (server) is responsible for getting the traversal times of the vehicles. From the traversal time, the centralized entity shows the results in different VANET scenarios. The proposed work also does not compute to the traffic congestion traffic on the road.

Ahmed et al. proposed a novel scheme called IVCD (infrastructure-based Vehicle Congestion Detection) to support vehicle congestion detection and speed estimation [5]. By using iterative COC (Context-Oriented Communication) information, the proposed IVCD extracts the protection period (time-headway) between vehicles. The IVCD mechanism detects traffic congestion between vehicles over time and disseminates the identification results locally as well as globally. Other well-known schemes are [6], [16], [17] proposed for traffic congestion detection. The purposed schemes studied the relationship between different traffic parameters like traffic density/congestion and vehicle's speed and showed how these parameters could affect each other. Additionally, the relation between congestion/density and speed is the most critical parameter in the purposed schemes. In [6], the authors suggested the following linear relation of speed-density, as follows:

$$\mu = \mu_{fs} - \left(\frac{\mu_{fs}}{D_j}\right) * D, \tag{1}$$

where free-speed signified with μ_{fs}, D shows the traffic density, μ shows mean speed at D, and jam-density represented with D_j, as shown in Eq. (1). The authors in [17], proposed a logarithmic speed-density relationship, as shown in Eq. (2), where μ_m represent the speed at maximum flows.

$$\mu = \mu_m \ln\left(\frac{D_j}{D}\right) \tag{2}$$

In [7], the authors proposed a probabilistic scheme to gather traffic data based on the V2I architecture. In the proposed scheme, two new techniques are used to detect accident events in the road traffic scenario automatically. During the accident event, a vehicle interacting with the RSU on the specified channel at regular intervals, which is available on the roadside. An approximation of the instantaneous congestion scheme of the vehicle is also proposed in [8], based on data obtained by the RSU and vehicle. This scheme calculates the congestion level from the RSU messages, and RSU collects the beacon messages from both architecture, such as V2V and V2I.

In this paper, we use SD-VANET architecture, unlike traditional VANET architecture and as well as all the works mentioned above, which distributes the traffic on different road proactively to mitigate the traffic congestion.

B. SD-VANET Architecture

Recently, SDN-based architecture has been proposed for VANET to solve network problems. This section explains the related work to SD-VANET architecture as follows:

In [11], the authors proposed an SD-VANET architecture that provides scalability and flexibility in different operational modes. In case of connectivity failure with SDN controller or Base Station (BS), the proposed work installs local SDN agents in every vehicle to optimize the performance of the network. Additionally, the results show that centralized architecture (i.e., SD-VANET) performs better than the traditional distributed architecture. However, the proposed work does not discuss the about traffic congestion management and feasible path. The dynamic nature of VANET, wireless links are vulnerable because of the high mobility of vehicles, which leads to packets loss. In dynamic VANET, link stability plays a vital role in increasing the packet delivery ratio. In [12], the authors proposed a novel routing scheme in SDVN (Software-Defined Vehicular Network) to forward the packets on multiple shortest paths.

In HetNets, to enable the communication between vehicles, the authors proposed an SDVN architecture [18]. The authors also explain the challenges in the SD-VANET architecture and highlight the opportunities of this integrated architecture. The proposed work focus on the heterogeneity problem in the VANET. Additionally, the proposed SDVN minimize the frequency of status update of the vehicles by using the trajectory predictions. They use POX as an SDN controller with an NS3 simulator to validate their proposed architecture performance. In [19], the authors using the SDN controller, to collect all the information about the vehicles in the network, like vehicle speed, direction, geographical position, and neighboring RSUs' ID using 802.11p. Based on the information, the proposed scheme takes the offloading and handover decision. The proposed scheme is called OHD-SDN (Offloading with Handover Decision based on SDN).

B. Dong et al. proposed on-demand routing in the SD-VANET architecture called SDAO [20]. The proposed SDAO architecture consists of two levels, centralized local level and distributed global level. The centralized local level computes a route for every vehicle, and for global routing distributed global level is responsible for reducing the route computation overhead in VANET. In [21], the authors proposed SD-VANET architecture to support the next generation (5G) communication. The proposed architecture also provides efficient resource utilization, flexible control, and network management. They also used fog computing architecture to minimize the delay, control overhead on the SDN controller, and also maximize the throughput. In [22], [23], [24], the authors also proposed energy-efficient routing for VANET architecture using SDN and fog computing architecture to decrease the control overhead on the SDN controller, and also maximize the throughput.

The above related work shows that mostly SD-VANET focused on the minimize the SDN controller overhead, maximize the throughput, minimize the delay, and high resource utilization. However, the traffic congestion problem is missing in related work. Additionally, traffic congestion leads to the vehicle's accident, increases the trip time, and prolong the delay. Thus, in the proposed SD-VANET architecture, we try to minimize the traffic congestion and provide a feasible alternative subject to the less end-to-end delay.

III. Theoretical Problem Explanation

The VANET (i.e., ITS) attracted considerable attention because it is now a phenomenon and provisioning a variety of new services like traffic safety, avoid traffic congestion, and enhance traffic flow. The ITS provides traffic alerts, mobile cloud services, route planning, and roadside safety [1], [2], [5]. However, some critical issues in the traditional VANET architecture in urban areas are still unresolved like traffic congestion, which leads to long queuing delay, longer trip times, and decrease the vehicle's speed [6], [7], [8]. The exponential increment in connected vehicles leads to the traffic congestion problem. Traffic congestion is one of the severe problems which can paralyze the complete VANET architecture. The VANET architecture, the devices not only represent the connection among the vehicles but also include communication among infrastructure, pedestrians' collaboration, and the roads. Based on the forecast, over 300 million vehicles are emerging into the VANET market in the coming years [25]. To reducing the traffic congestion problem as mentioned earlier, it would be imperative to inform the vehicles timely to select the feasible alternative route toward destination.

To clarify the above problem statement, we consider a network, as shown in Fig. 3. In this scenario, we consider two different destination points (A and B); for type-A vehicles, the destination point is A, and for type-B vehicles, the destination is B. There are multiple paths for both destinations (A and B) in the given Fig. 1, but for the sake of simplicity, we consider four (4) paths for type-A vehicles, and type-B vehicles, we assume three (3) paths (see Fig. 3). For example, all the vehicles (type-A) select the path based on delay (i.e., path-1). This produces congestion on the path-1 and prolong the queuing delay and decrease vehicle speed. The same case with type-B vehicles, if they select the path based on delay (i.e., Path-1 or Path-2), this produces traffic congestion. The traditional VANET routing protocol provides the shortest path to the vehicles (i.e., ITS), which may lead to the congestion problem. As mentioned earlier in the problem statement that traffic congestion not only prolong the queuing delay, longer trip times, and decrease vehicle speed [5], [6] and paralyze the traffic system, but also waste the time of travelers. Therefore, it is essential to detect traffic congestion and rapid action accordingly.

To handle the traffic congestion problem in VANET, we should compute the route or divert the traffic based on the number of vehicles (vehicles' density) on route to minimize the congestion problem and minimize the delay. For this, we need a global view of the network to handle the congestion problem in VANET. Therefore, SD-VANET architecture possibly plays a vital role in reducing road traffic congestion because of the congestion problem prolong trip time, increase queuing delay, and also decrease the vehicle's speed. As mentioned earlier, the logical centralized SDN architecture decouples the control plane (controller) for data plane devices (forwarding devices) like routers, switches, APs, and RSU. Thus, SDN makes the forwarding devices (i.e., routers/switches/RSU) programmable [26]. These forwarding devices send information to the control plane, also known as the SDN controller. This, in turn, makes it easy to manage and to control the VANET network.

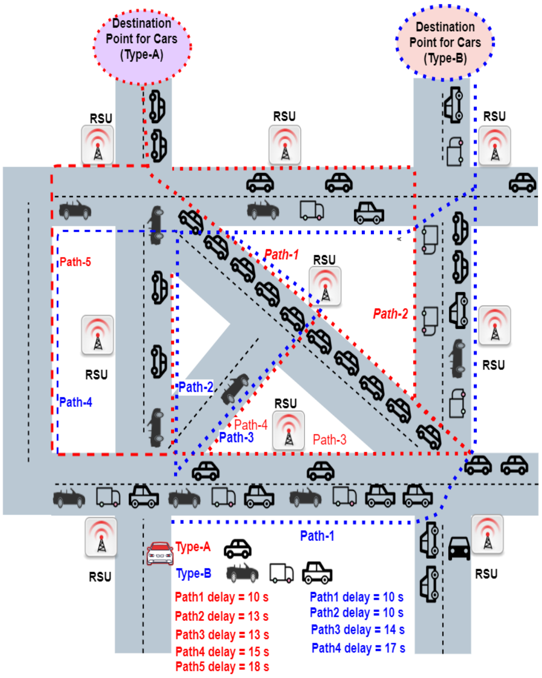


Fig. 3. Congestion scenario in VANET Architecture (shortest Path Selection).

The SDN controller continuously collects information from the OpenFlow enabled RSUs about the entire network to make intelligent decisions about traffic control. In SD-VANET architecture, in wireless medium exchange beacon messages (a standard message in the VANET architecture, to lean information about the neighbor's vehicles) periodically to collect the information the network. The logical centralized controller uses this information to create a network graph. Additionally, using this information, the SDN controller can handle the congestion problem proactively distribute the vehicles according to vehicles' density on the roads and provide a feasible path and less congested path in the network. OpenFlow protocol is uses to collect the information from the OpenFlow-enabled devices. When the SDN controller receives a new request for the vehicle, then search for less congestion path subject to delay parameter. After the path computing, the SDN controller informs the vehicle about the less congested path. For the sake of simplicity, consider Type-A vehicles. Four (4) paths are available toward destination-A. If the SDN controller distributes the vehicles on different paths, noticeable, it can decrease the congestion problem and minimize the queuing delay.

IV. Problem Formulation

This section articulates the articulate the problem statement (as describe in Section III) to handle the challenging task to compute the best path (i.e., a congestion-free path and average minimum end-to-end) from the source vehicle to the chosen destination. Our proposed heuristic algorithms Congestion Free Path (CFP) and OCFP proactively disturbed the vehicles in the SD-VANET architecture to minimize traffic congestion and also decrease the average delay. For quick reference, first, we summarize the significant notations in Table I.

SDN framework primarily designed for wired networks, but now it is widely used in wireless and mobile networks [11], [18]. The SDN framework offered centralized control to optimize the resources in the wireless network like channel allocation, congestion avoidance, interference avoidance.

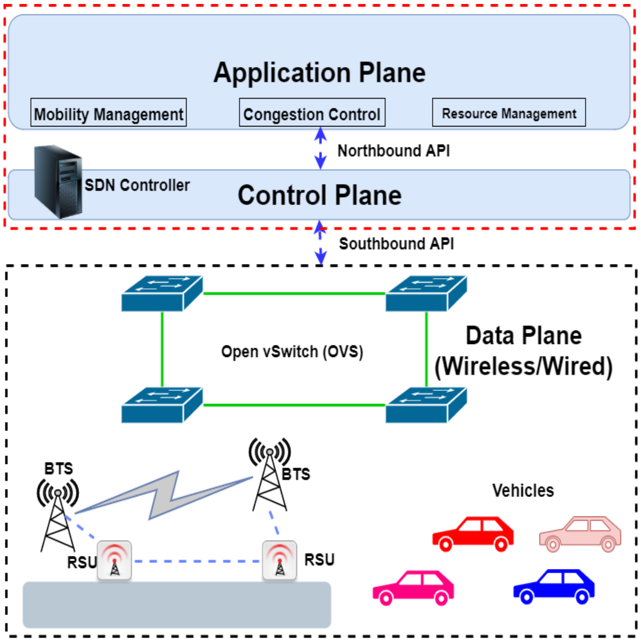


Fig. 4. SD-VANET Framework.

TABLE I. NOTATION DEFINITION

Notation	Definition
G	road network signified as a directed graph
V	set of nodes, such as points of interest or vertices, where $v_i \in V$
R	set of roads (i.e., links), where $r \in R$
r_c	road capacity (the maximum rate at which vehicles can travel on the road during a given time)
r_t	time spent to traverse a road (seconds (s))
r_l	the traffic load on the road (number of vehicles)
r_q	delay parameter (s)
λ_r	the cost function, calculated as the traffic load on the road r_l subjected to the delay parameter r_q
N	$n \in N$, where $N \subseteq V \times V$
(s_n, d_n)	pair of the source node and destination node
χ_n	traffic demand associated with n
$In(v_i)$	incoming roads
$Ou(v_i)$	outgoing roads
$T\chi_N$	total traffic demand between source and destination pairs (i.e., $(s_n,\ d_n)$)
ψ	maximum utilization of a road
r_μ	average end-to-end delay (s)

Packet forwarding in different scenarios (multi-path, multi-hop), efficiently handle the mobility issues, manage the heterogeneity and dynamic VANET environment. In SD-VANET, the control plane (controller) is responsible for possesses all the logical functionalities and takes actions on behalf of data plane devices, like routing, topology control, congestion avoidance, and mobility management, as shown in Fig. 4. The data plane consists of different types of devices or nodes like RSU, vehicles, OBUs (On-Board Units), and OpenFlow switches. These data plane devices execute actions specified by the flow rules which are supplied by the controller (control plane) [11], [12], [18]. Usually, for communication, the data plane devices such as OBU and RSU use wireless interfaces such as DSRC (Dedicated Short-Range Communication) and LTE.

To avoid traffic congestion in SD-VANET, the controller periodically collect the network information from OpenFlow enabled RUS [11], [12]. The controller uses this information to build a global network (G) to make intelligent decisions, such as routing, mitigation of congestion, and mobility management. The traditional VANET finds the path based on the collective functioning of devices; however, in SD-VANET, the path computation relies on the logically centralized SDN controller. The fact, the SDN controller, maintains a global network view and it simplified the congestion management in SD-VANET.

In the proposed model, the road network signified as a directed graph $G(V, R)$. In G, V denotes a set of nodes, such as points of interest or vertices, where R is the set of roads (i.e., links). In G, each road $r \in R$ associated road capacity $r_c > 0$ and a time $r_t > 0$ spent to traverse the uncongested/unloaded road. For computing the congestion level of all roads in a network, the proposed CFP and OCFP algorithms use a cost function, signified as λ_r is calculated as the traffic load on the road r_l subjected to the delay parameter r_q. Consequently, $n \in N$ is the set source and destination pairs (i.e., (s_n, d_n)), where $N \subseteq V \times V$. Each source node s_n has associated with traffic demand χ_n towards the destination node d_n. In the road network, each node $v_i \in V$ have a set of incoming roads (i.e., links) and a set of outgoing roads, signified by $In(v_i)$ and $Ou(v_i)$ respectively. The total traffic demand χ_N can be $T\chi_N = \sum_{n \in N} \chi_n$.

Objective Function

The foremost objective of our proposed algorithms is to minimize the traffic congestion problem and the average delay in VANET architecture. More precisely, to minimize the maximum utilization of the roads and provide an average end-to-end delay, as shown in Eq. (3) and Eq. (4). The objective function Eq. (3) distribute the traffic (minimize the maximum utilization) among all roads. For an average end-to-end delay, the cost function for a road $r \in R$ is λ_r= $r_l r_t \left[1 + \left(\frac{r_l}{r_c}\right)^{r_q}\right]$. In real-world VANETs, road delay r_q is commonly defined by nonlinear function subject to the congestion parameter. Here, we assume that the cost function λ_r is an increasing function. In Eq. (4), we normalized the average end-to-end delay r_μ by dividing $T\chi_N$. Furthermore, it shows the end-to-end delay for the traffic on each road is minimized. The constraints of Eq. (3) and Eq. (4) are explained as follows. The maximum utilization ψ of a road, as shown in constraint Eq. (5). The total traffic χ_n of all source nodes on a road $r \in R$, is shown in the constraint Eq. (6) with the decision variable ∂_r^n, where ∂_r^n is the proportion of commodity (i.e., (s_n, d_n)) $n \in N$ on a road r. The traffic conservation is shown in constraint Eq. (7), and constraint Eq. (8) defines the domain of decision variables. The objective function provides vital statistics about the traffic and distributes the traffic efficiently if $\psi < 1$ (i.e., without beyond the road capacity level $\left(\frac{r_l}{r_c} \leq \psi\right)$.

Decision Variables

maximum road utilization = ψ
traffic load on a road = r_l
total traffic ((proportion of commodity, (i.e., (s_n, d_n))
$n \in N$ *on the road* r = ∂_r^n

$$Objective1 \quad min \quad \psi \tag{3}$$

$$Objective2 \;\; min \;\; r_\mu = \sum_{r \in R} r_l r_t \left[1 + \left(\frac{r_l}{r_c} \right)^{r_q} \right] / T_{\chi_N} \qquad (4)$$

$$\frac{r_l}{r_c} \leq \psi, \quad \forall \; r \in R \qquad (5)$$

$$r_l = \sum_{n \in N} \chi_n \partial_r^n, \quad \forall \;\; r \in R \qquad (6)$$

$$\sum_{r \in In(v_i)} \partial_r^n - \sum_{r \in Ou(v_i)} \partial_r^n = \begin{cases} 1, & if \;\; v_i = d_n \\ -1, & if \;\; v_i = s_n \\ 0, & otherwise \end{cases} \quad \forall r \in R, n \in N \qquad (7)$$

$$\partial_r^n \in [0, 1], r_l \geq 0, \forall r \in R, \forall n \in N, \psi \leq 1 \qquad (8)$$

The SDN controller continuously collects information about the entire network to make intelligent decisions about traffic control. The proposed SD-VANET framework, in wireless medium exchange beacon messages (a standard message in the VANET framework, to lean information about the neighbor's vehicles) periodically to collect the information the network [11]. The SDN controller uses this information to create a network connectivity graph G. Additionally, using this information, the SDN controller can handle the congestion problem and provide a feasible path in the network. When the SDN controller receives a new request, then search for less congestion path subject to delay parameter, based on Eq. (3) and Eq. (4) subject to the mentioned constraints. After the path computing, the SDN controller informs the vehicle about the less congested path, as shown in Algorithm 1. When the SDN controller receives a path request from a new vehicle, if the requested path is congestion-free (i.e., $\psi \leq 0.5$), then the SDN controller returns the path (see Step 2). Otherwise, the SDN controller checks the alternative path for the vehicle (see Step 3). In Algorithm 1, the SDN controller searches all the alternative routes and return the optimum route based on Eq. (3) and Eq. (4) subject to additional constraints. However, searching for all routes increases the computational time and also NP-hard. Thus, to minimize the computational time, we proposed, the proposed Optimized heuristic CFP algorithm (OCFP) (see Algorithm 2), the controller only computes the best path among the "k" (i.e., "k" = 15) alternative routes. This is an optimization problem; therefore, to minimize the computational time, the proposed OCFP algorithm only searches for the "k" path based on Eq. (3) and Eq. (4) subject to mentioned constraints.

Algorithm 1: CFP in SD-VANET

```
input : Graph of road map $G(V, R)$
output: reliable congestion free path subject to the
        minimum utilization $\psi$ and average minimum
        delay $r_\mu$
1 $SDN\ Controller \leftarrow$ path_request (new_vehicle)
2 $if \;\; \psi \leq 0.5:$
      the requested path is congestion-free
      $return \;\; path\_request$
3 $else \;\; :$
      (i) check all alternative congestion-free paths
      (ii) select path based on $\psi$ and $r_\mu$, subject
           to the constraints
      (iii) return path_request (update the vehicle
           about the alternative path)
4 $update\ Graph\ road\ map\ status\ (G\ (V, R))$
5 $end$
```

Algorithm 2: OCFP in SD-VANET

```
input : Graph of road map $G(V, R)$
output: reliable congestion free path subject to the
        minimum utilization $\psi$ and average minimum
        delay $r_\mu$
1 $SDN\ Controller \leftarrow$ path_request (new_vehicle)
2 $if \;\; \psi \leq 0.5:$
      the requested path is congestion-free
      $return \;\; path\_request$
3 $else \;\; :$
      (i) check "k" alternative congestion-free paths
           (i.e., "k" = 15)
      (ii) select path based on $\psi$ and $r_\mu$, subject
           to the constraints
      (iii) return path_request (update the vehicle
           about the alternative path)
4 $update\ Graph\ road\ map\ status\ (G\ (V, R))$
5 $end$
```

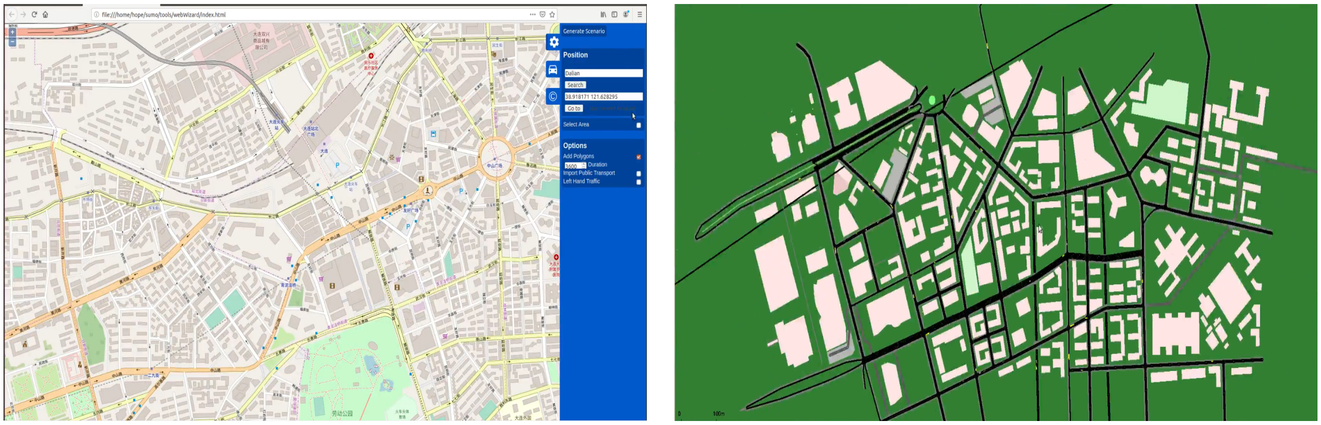
V. Simulation Setup

This section presents the simulation and configuration setups in various scenarios to validate the proposed algorithms in SD-VANET architecture. We use NS3[1] simulator for the model architecture, and for the urban road network, we used SUMO[2] to generate the real vehicle mobility traffic of Dalian city, China using OpenStreetMap (OSM). OpenStreetMap shows a layout of the road network, as shown in Fig. 5. Geographically, every road network contains the number of alternative paths, but traditional VANET routing protocols are distributed in nature and provide the shortest path. The shortest path causes the congestion problem and decreases the reliability of the network like prolong delay, increases trip time. The foremost objective of the proposed heuristic algorithms is to minimize traffic congestion and also provide a feasible path to a vehicle subject to the less average delay.

Thus, our goal is to minimize the problem of traffic congestion in urban areas; we generate a road traffic scenario of Dalian city, China, one square kilometer $(1km * 1km)$ in scales. The scenario for road traffic consists of six-vertical and six-horizontal roads, with junction/intersection every 250 meters. Each road is one kilometer long and two meters wide with two (2) lanes in each direction. Subsequently, at junctions, vehicles can move straight or can turn left or right. We kept constant the total number of vehicles, that is to say, 500. In vehicles, car probability is 0.80, and bus probability is 0.20, while car length is 5m, and bus length is 10m. A vehicle's maximum speed is set at 40 km/h; however, the vehicle's speed changes with time because the SUMO also simulates the traffic lights as well. The vehicles grouped in 10 traffic flows. Each flow is taking a different path/route in a total distance. At junctions, these paths intersect to simulate a high number of vehicles on the road (i.e., more than road capacity).

[1] NS3 Homepage, https://www.nsnam.org/

[2] SUMO Homepage, http://sumo.sourceforge.net/



(a) OpenStreetMap of Dalian City

(b) SUMO Network View

Fig. 5. (a) Map of Dalian city obtain from OpenStreetMap. (b) SUMO Network View to OSM of Dalian road network

This would turn in traffic congestion, which our algorithms are intended to mitigate. The total simulation time is 500s. We simulate the proposed algorithms in NS-3 and measure its performance against SDN-based Shortest Path (SDN-based SP) and an existing Distributed Road Traffic Congestion (DRTC) [6].

A. Performance Metrics

To validate the performance of the proposed algorithms, we examine the following performance parameters are considered:

- Monitor the Vehicle's Speed: We monitor the vehicle's speed by varying the congestion parameter, i.e., $\psi \leq 0.5$, $\psi \leq 0.7$, and $\psi \geq 1$.
- Congestion ratio: Congestion ratio means the density of vehicles on each path. In the simulation, a road can be considered as a congested road, if $\psi \geq 1$.
- Underutilized Roads: In the simulation, we examine the utilization ratio of roads in the network. In the simulation, a road can be considered as an underutilized road, if $\psi < 0.4$.

B. Experimental Results and Analysis

In this section, we would like to show the performance of our proposed congestion mitigation algorithms, i.e., CFP, OCFP, compared to SDN based Shortest Path (SP), and an existing Distributed Road Traffic Congestion (DRTC) [6]. In particular, we evaluate the impact number of vehicles on the congestion level. Moreover, we compare the performance in terms of vehicle's speed by varying the congestion parameter, congestion level, and underutilized roads.

To evaluate the proposed scheme (i.e., CFP), we exam the vehicle's speed over different congestion levels, i.e., $\psi \leq 0.5$, $\psi \leq 0.7$, and $\psi \geq 1$. In the simulation, the maximum speed of a vehicle is 40km/h (11.11 m/s), but the actual speed is based on the traffic conditions like congestion and traffic lights. Therefore, we evaluate the vehicle speed by varying traffic congestion level. Consequently, we selected one vehicle and plotted its speed under different congestion levels, as shown in Fig. 6, Fig. 7, and Fig. 8. From these results, we exam the congestion level significantly affect the vehicle's speed. In Fig. 6, the congestion level is $\psi \leq 0.5$; therefore, the vehicle's speed is affected less compared to Fig. 7 and Fig. 8. When the congestion level reaches the maximum level (i.e., $\psi \geq 1$), then it affects the vehicle's speed dramatically, as shown in Fig. 8.

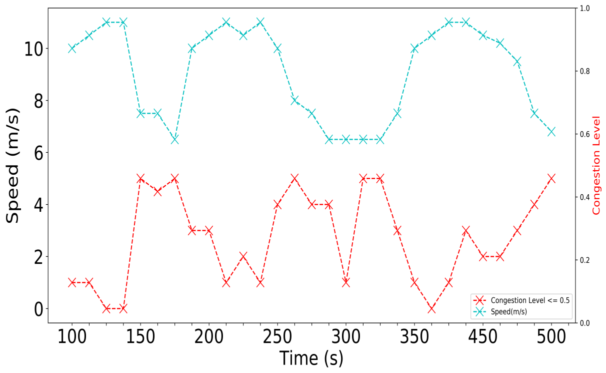


Fig. 6. Vehicle's Speed vs. Congestion Level ($\psi \leq 0.5$).

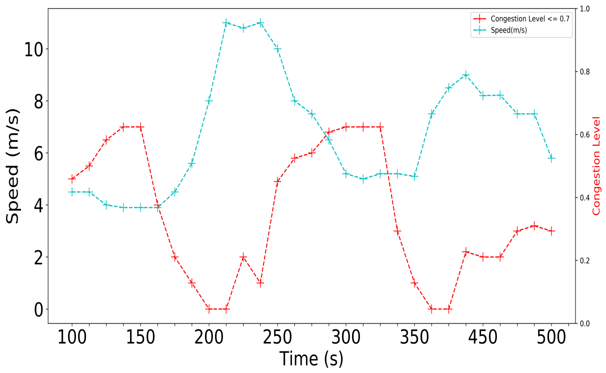


Fig. 7. Vehicle's Speed vs. Congestion Level ($\psi \leq 0.7$).

The result shows that the congestion level increases in all schemes when the number of vehicles increases, as shown in Fig. 9. The result shows that in our proposed approaches (i.e., CFP and OCFP), the traffic congestion level is very less compared to SDN-based SP and DRTC approaches. The congestion

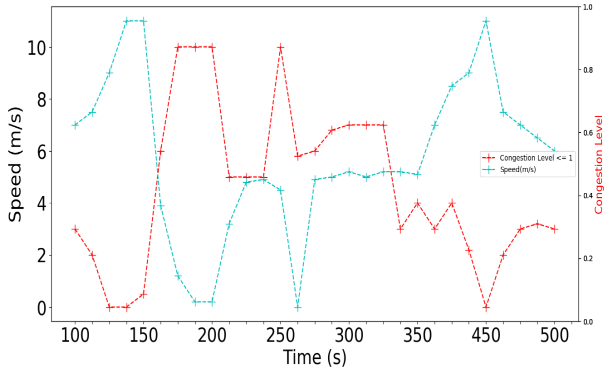


Fig. 8. Vehicle's Speed vs. Congestion Level ($\psi \leq 1$).

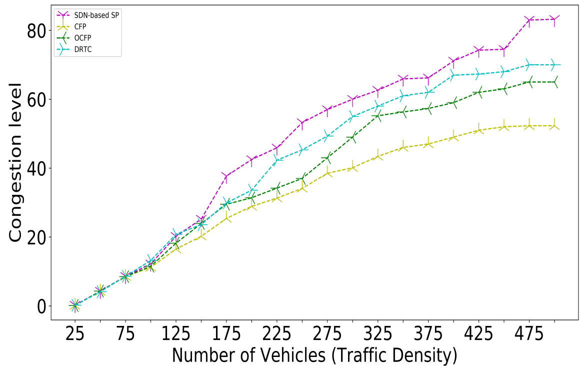


Fig. 9. Congestion Level Vs. Traffic Density.

level is almost the same in all approaches when the traffic density is between 25 to 125. However, when the traffic density increases, the proposed approaches perform better. The main reason is that the proposed approaches distribute the vehicles proactively on the alternative paths according to the vehicles' density on the road. The congestion level is very less in CFP approach because of this scheme's search for all the alternative paths for a vehicle. Additionally, the congestion level in OCFP is more than CFP because this scheme only searches for the "k" alternative path based on the objective function, unlike CFP. Subsequently, SDN-based SP finds the shortest path and does not consider the vehicles' density on the road, which leads to traffic congestion. Furthermore, the congestion level in DRTC is more compared to proposed algorithms because DRTC is an infrastructure-less distributed V2V scheme for congestion detection. The DRTC scheme enables each vehicle to detect the traffic congestion condition and then share the congestion information with vehicles through cooperation. The cooperation, however, decreases the number of broadcasting vehicles but lack of a centralized controller, the congestion level is more than the proposed algorithms. Additionally, the DRTC scheme does not distribute the vehicles proactively on different paths. The congestion level in CFP is almost 53%, 65% in OCFP, in DRTC, the congestion level is almost 70%, and in SDN-based SP, the congestion level is more than 80% when the traffic's density reached to 500.

Fig. 10 shows the result of the underutilized ratio of roads in VANET architecture. The underutilized ratio of roads decreases in all schemes when the number of vehicles increases.

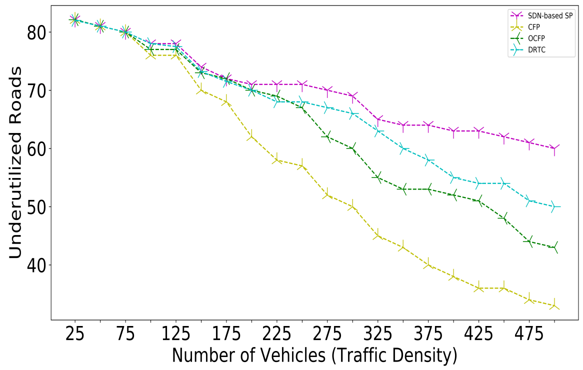


Fig. 10. Underutilized Roads Vs. Traffic Density.

However, the result shows that in our proposed approaches (i.e., CFP and OCFP), distribute the vehicles proactively on all paths and utilize the more roads compared to SDN-based SP and DRTC approaches. Therefore, the underutilized ratio of roads less in the proposed approach. The utilization ratio of roads in all approaches is almost the same when the traffic density is between 25 to 125. However, when the traffic density increases, the proposed approaches perform impressively because the proposed approaches distribute the vehicles based on traffic density on the roads. The underutilized ratio is very less in CFP approach because of this scheme's search for all the alternative paths for a vehicle, as explained in Algorithm 1. Additionally, the underutilized ratio in OCFP is more than CFP because this scheme only searches for the "k" alternative path based on the objective function, unlike CFP. The SDN-based SP and DRTC schemes do not distribute the vehicles on all roads; therefore, the underutilized ratio is more than the proposed approaches. The underutilized ratio in CFP is almost 32%, 43% in the OCFP scheme, in DRTC approach, the congestion level is almost 50%, and in SDN-based SP, the congestion level is more than 60% when the traffic's density reached to 500.

The results of all scenarios indicate that our proposed algorithms, CFP and OCFP, outperform compared to DRTC. The primary difference between our proposed algorithms and DRTC is that our algorithms use centralized SDN architecture, and DRTC uses distributed architecture. Secondly, DRTC uses Eq. (1) to compute the congestion level; however, our algorithm distributed the vehicles proactively on alternative paths. Thus, this is the main reason that the proposed algorithms have global information about the network and provide better services. The proposed algorithms decrease the congestion problem in VANET.

VI. Conclusion

The proposed algorithms try to provide the congestion-free path subject to the average minimum delay paths to the vehicles in the SD-VANET network. In VANET, the congestion problem leads to longer trip times, decreases the vehicle's speed, and prolong delay. The SDN controller in VANET collects the information about the vehicles on each road to calculate the traffic density on the road. If the road is going to dense, the SDN controller uses global network information and divert the traffic on another feasible path subject to minimum delay. The results indicate that the proposed algorithms (i.e., CFP and OCFP) in SD-VANET architecture effectively switches the vehicle on an un-congested path subject to minimum delay and minimize the congestion characteristics. Our proposed approach is a heuristic approach and has linear time complexity. For future work, we will use some machine learning approaches to minimize the time complexity and provide better performance.

REFERENCES

[1] A. Wahid, A. Rao, and D. Goel, "Server communication reduction for gps-based floating car data traffic congestion detection method," in *Integrated Intelligent Computing, Communication and Security*. Springer, 2019, pp. 415–425.

[2] Z. Zhou, H. Yu, C. Xu, Y. Zhang, S. Mumtaz, and J. Rodriguez, "Dependable content distribution in d2d-based cooperative vehicular networks: A big data-integrated coalition game approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 953–964, 2018.

[3] X. Ke, L. Shi, W. Guo, and D. Chen, "Multi-dimensional traffic congestion detection based on fusion of visual features and convolutional neural network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 6, pp. 2157–2170, 2018.

[4] C. Jayapal and S. S. Roy, "Road traffic congestion management using vanet," in *2016 International Conference on Advances in Human Machine Interaction (HMI)*. IEEE, 2016, pp. 1–7.

[5] M. Ahmad, Q. Chen, Z. Khan, M. Ahmad, and F. Khurshid, "Infrastructure-based vehicular congestion detection scheme for v2i," *International Journal of Communication Systems*, vol. 32, no. 3, p. e3877, 2019.

[6] M. Milojevic and V. Rakocevic, "Distributed road traffic congestion quantification using cooperative vanets," in *2014 13th annual Mediterranean Ad Hoc networking workshop (MED-HOC-NET)*. IEEE, 2014, pp. 203–210.

[7] O. Popescu, S. Sha-Mohammad, H. Abdel-Wahab, D. C. Popescu, and S. El-Tawab, "Automatic incident detection in intelligent transportation systems using aggregation of traffic parameters collected through v2i communications," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 2, pp. 64–75, 2017.

[8] J. Barrachina, P. Garrido, M. Fogue, F. J. Martinez, J.-C. Cano, C. T. Calafate, and P. Manzoni, "A v2i-based real-time traffic density estimation system in urban scenarios," *Wireless Personal Communications*, vol. 83, no. 1, pp. 259–280, 2015.

[9] M. Dixit, R. Kumar, and A. K. Sagar, "Vanet: Architectures, research issues, routing protocols, and its applications," in *2016 International Conference on Computing, Communication and Automation (ICCCA)*. IEEE, 2016, pp. 555–561.

[10] T. Nadeem, S. Dashtinezhad, C. Liao, and L. Iftode, "Trafficview: A scalable traffic monitoring system," in *IEEE International Conference on Mobile Data Management, 2004. Proceedings. 2004*. IEEE, 2004, pp. 13–26.

[11] I. Ku, Y. Lu, M. Gerla, R. L. Gomes, F. Ongaro, and E. Cerqueira, "Towards software-defined vanet: Architecture and services," in *2014 13th annual Mediterranean ad hoc networking workshop (MED-HOC-NET)*. IEEE, 2014, pp. 103–110.

[12] K. L. K. Sudheera, M. Ma, and P. H. J. Chong, "Link stability based optimized routing framework for software defined vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 2934–2945, 2019.

[13] L. Zhao, W. Zhao, A. Al-Dubai, and G. Min, "A novel adaptive routing and switching scheme for software-defined vehicular networks," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–6.

[14] L. Wischhof, A. Ebner, and H. Rohling, "Information dissemination in self-organizing intervehicle networks," *IEEE Transactions on intelligent transportation systems*, vol. 6, no. 1, pp. 90–101, 2005.

[15] G. Marfia and M. Roccetti, "Vehicular congestion detection and short-term forecasting: a new model with results," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 7, pp. 2936–2948, 2011.

[16] M. Kimura, Y. Taoda, Y. Kakuda, S. Inoue, and T. Dohi, "A novel method based on vanet for alleviating traffic congestion in urban transportations," in *2013 IEEE Eleventh International Symposium on Autonomous Decentralized Systems (ISADS)*. IEEE, 2013, pp. 1–7.

[17] Z. He, J. Cao, and T. Li, "Mice: A real-time traffic estimation based vehicular path planning solution using vanets," in *2012 International Conference on Connected Vehicles and Expo (ICCVE)*. IEEE, 2012, pp. 172–178.

[18] Z. He, J. Cao, and X. Liu, "Sdvn: Enabling rapid network innovation for heterogeneous vehicular communication," *IEEE network*, vol. 30, no. 4, pp. 10–15, 2016.

[19] C.-M. Huang, M.-S. Chiang, D.-T. Dao, H.-M. Pai, S. Xu, and H. Zhou, "Vehicle-to-infrastructure (v2i) offloading from cellular network to 802.11 p wi-fi network based on the software-defined network (sdn) architecture," *Vehicular Communications*, vol. 9, pp. 288–300, 2017.

[20] B. Dong, W. Wu, Z. Yang, and J. Li, "Software defined networking based on-demand routing protocol in vehicle ad hoc networks," in *2016 12th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN)*. IEEE, 2016, pp. 207–213.

[21] A. A. Khan, M. Abolhasan, and W. Ni, "5g next generation vanets using sdn and fog computing framework," in *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2018, pp. 1–6.

[22] A. J. Kadhim and S. A. H. Seno, "Energy-efficient multicast routing protocol based on sdn and fog computing for vehicular networks," *Ad Hoc Networks*, vol. 84, pp. 68–81, 2019.

[23] J. C. Nobre, A. M. de Souza, D. Rosário, C. Both, L. A. Villas, E. Cerqueira, T. Braun, and M. Gerla, "Vehicular software-defined networking and fog computing: Integration and design principles," *Ad Hoc Networks*, vol. 82, pp. 172–181, 2019.

[24] N. B. Truong, G. M. Lee, and Y. Ghamri-Doudane, "Software defined networking-based vehicular adhoc network with fog computing," in *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*. IEEE, 2015, pp. 1202–1207.

[25] K. Liu, X. Xu, M. Chen, B. Liu, L. Wu, and V. C. Lee, "A hierarchical architecture for the future internet of vehicles," *IEEE Communications Magazine*, vol. 57, no. 7, pp. 41–47, 2019.

[26] J. Bhatia, R. Dave, H. Bhayani, S. Tanwar, and A. Nayyar, "Sdn-based real-time urban traffic analysis in vanet environment," *Computer Communications*, vol. 149, pp. 162–175, 2020.

Perceived Usability of Educational Chemistry Game Gathered via CSUQ Usability Testing in Indonesian High School Students

Herman Tolle [1], Muhammad Hafis[2]
Ahmad Afif Supianto[3]
Faculty of Computer Science
Brawijaya University, Malang, East Java, Indonesia

Kohei Arai[4]
Graduate School of Science and Engineering
Saga University
Saga, Japan

***Abstract*—Educational game is now a commonplace among students and teachers alike. Recent researches show that studies regarding educational game general effectiveness in the learning environment are nothing new. However, usability studies in the educational game are rather rare compared to general non-game-related usability studies. This research synthesizes the result obtained from the Computer System Usability Questionnaire (CSUQ) and separated between multiple students pre-existing grouping such as genders, prior knowledge, as well as experimental treatment setup such as materials given before the game session. The metrics are tested in an Indonesian high school by using an educational game of chemistry regarding the topic of reaction rate with a total of 53 participants. General results show that there exist many differences of perceived usability aspects between male and female students, the existence of learning materials given before the game session, as well as the existence of students' prior knowledge. Overall, the main findings of this research show that usability in the educational game is affected by gender, materials existence, and previous knowledge existence.**

Keywords*—Usability testing; CSUQ; educational game; male students; female students

I. Introduction

In recent years, the digital educational game has emerged as one of the more sophisticated methods to augment the student learning process. With the increasing ease of access to technology [1], students are more exposed to computers and smartphones. They use it more than ever as the digital educational game approach can improve students' motivation; as such, the field of technology-enhanced learning is now getting more important than ever [2]. Some clear advantage of educational games is how the students perceive it to be useful for their learning experience. Recent studies show that educational game is perceived to be able to increase students' enjoyment during the learning process [3] as well as promoting skill and knowledge gain [4]. Digital educational games also offer an advantage in terms of enriched visuals as well as more appealing multimedia aspects [5]. In terms of subjects, multiple domains of educational topics have been adapted in the form of digital educational games such as art [6], language learning [7], and mathematics [8]. The current state-of-the-art of digital educational game shows that it is an emerging approach to supplement conventional generally-used instruction-based approach.

Generally, the essential aspect of a digital educational game is whether the game can enable students' knowledge acquisition. This is generally done by evaluating students' performance in a quasi-experimental setup and aimed to evaluate the effectiveness of the learning environment [9]. However, the way students as end-users interact with the game itself also plays a significant role in ensuring optimal knowledge acquisition process [10], as usability is positively strongly correlated with increased learning motivation [11]. Similar to other software, digital educational games also require proper quality assurance in terms of usability.

Usability is a broad term defined as "user-friendliness" of software and quality that attributes the ease-of-access of an interface in a software [12]. Usability in the digital educational game, however, focuses on ensuring students learn effectively and efficiently as well as maintaining students' interest in the game itself [13]. In terms of how general usability is being measured, there exist several approaches to acquire different information regarding usabilities, such as observational technique [10] and think-aloud technique [14]. Different metrics to quantify a different aspect of usability also exist and is used to quantify a different aspect of perceived usability such as System Usability Scale (SUS) [15] Usability Metric for User Experience (UMUX) [16], and Computer System Usability Questionnaire (CSUQ) [17]. Each metrics has a different purpose and assesses a different aspect of usability.

Recent existing research about usability categorizes the test subjects based on multiple classifications. Game-dependent skill-based classification usability has been done to detect whether there exists any difference in perceived usability between the classes [18]. A more general gender-based classification for usability testing has also been done before [13]. Regardless, the existing researches separate the usability criteria differently.

Specifically speaking, existing research focuses on using a particular metric for evaluation purposes [18]; however, in-depth research regarding each aspect and category of usability in a particular metric is also needed. A synthesis of information based on usability scores can gather critical aspects of the users when viewed from different demography and classifications [19]. An in-depth usability study is able to analyze users' satisfaction and create a recommendation for system improvement in the future [20]. Moreover, an existing in-depth study is done for general software and systems, but not for digital educational games. The urgency of an in-depth usability analysis from the digital educational game perspective is needed since digital educational games, and general software is vastly different. The pedagogical aspect and the delicate nature of students compared to general users should be taken as a primary consideration compared to only general usability aspect.

Another perspective is how the educational game is being deployed to the students. Also, the different pre-existing conditions of the students themselves, such as its skill level and current grade [18] or its treatment during experimental setup [21], have to be considered. Different treatment may result in a different result, either from students' study performance results or its usability.

Recent studies and development have been done on a digital educational game for high school students focusing on the subject of chemistry, from its design phase [22] and its performance based on students' scores [23]. The result shows that the digital educational game is effective at improving students' knowledge acquisition process. However, the usability aspect of the developed digital educational game has not been analyzed in-depth. This paper aims to synthesize the gathered usability test result by using one of the existing metrics for usability (CSUQ) and analyze the result based on its end-users (students) details during the game's experimental setup treatments.

The paper is then organized as follows. After the introduction, the second section will cover some theoretical background and related works, specifically the ones related to digital educational games as well as CSUQ itself. The third section introduces the developed digital educational game and its basic mechanics. The fourth section will cover the experimental setup. The fifth section will cover the results and discussion. Finally, the sixth section concludes the paper and discusses some future works.

II. Theoretical Background and Related Works

In this section, several reviews, and theoretical background related to digital educational games, usability, the Computer System Usability Questionnaire (CSUQ) as well as existing usability studies for educational games are presented.

A. Digital Educational Game

The very definition of digital educational games is quite hard to pin down, since there are several terms related closely with digital educational games, such as *gamification*, *Game-Based Learning* (GBL) as well as the popular educational game itself. Before defining digital educational games, to clear up the taxonomy, closely related terms are explained first. Gamification definition can be simplified as "the use of video game elements in non-gaming systems to improve user experience" [24] in which a gamified system mostly classified as a non-game system. GBL ramps up the usage of game elements for educational purpose, instead of just using some elements, GBL incorporate the game as an instructional strategy [25]. However, GBL does not necessarily mean that the adapted strategy is in a digital form. A narrower and more specific term related to GBL in regards to technology integration is *Digital Game-Based Learning* (DGBL) [26], which combines curricular contents and digital games to increase students' motivation [27]. With both related terms clearly defined to reduce confusion, the educational game can be finally defined.

Generally, the educational game can be defined as a game being designed and used for teaching and learning. Furthermore, it is also designed to help people to learn about a particular subject [28]. Compared to both gamification and GBL (or DGBL), educational game are much more focused on combining entertainment and learning in which the players do not feel like they are learning as in the conventional definition of learning [29]. However, the abovementioned definition of an educational game does not strictly define whether the educational game is in a digitized system. A more specific term for a digitized system for educational purposes is Digital Educational Games (DEG), in which the educational game is deployed in the form of software and has the purpose of teaching a particular subject [30]. Hence, a DEG is different from its related terms, whether from its definition, aspect of gaming being used, as well as how such aspect is being used.

B. Usability and Computer System Usability Questionnaire

Based on the ISO standard 9241-11 [31], usability is defined as to which extent a product can be used to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use. Usability in computer systems is used to learn users' experience (UX) when using an application. Usability is also used to measure the general usefulness of a system. There exist multiple standardized questionnaires that can be used to measure perceived usability [32]; one of them is the Computer System Usability Questionnaire (CSUQ). CSUQ is an instrument to measure perceived usability, which was developed by IBM [17] consisting of 19 validated questions using a 7-point Likert scale with an alpha score of 0.89, which represents a high degree of reliability. The questions are shown in Table I shown below with questions 1 through 8 representing the system usefulness, questions 9 through 15 representing the information quality section, questions 16 through 18 representing interface quality, and the overall satisfaction represented in question 19.

C. Usability Studies in Educational Games

Usability studies in educational games are rather scarce. A recent study has been done to analyze student's difficulty in playing a particular game by comparing the results between genders [33] with a result of male students performing slightly better compared to female students. Even in a non-game perspective, there exists a difference in preference between male and female students [34]. Another usability evaluation also has been done on an educational game for visually impaired users [35] in which the results show a promising potency for educational purposes. Both types of research however do not use the aforementioned CSUQ metrics but instead uses a personalized questionnaire. Evidence shows that some research has adapted the CSUQ for usability studies in an educational game [36] to acquire students' opinions towards a game. Another research also has been done with a modified CSUQ to increase relevance [37]. Generally, usability studies done in an educational context are done to test perceive students' perception towards an educational game, and the result of the studies is often used to improve the game. However, a further analysis of preexisting demographics and details of the students are often left out and not being used as a consideration to do such future improvements.

TABLE. I. CSUQ QUESTIONS BY CATEGORY

Category	Number	Question
System Usefulness (SYSUSE)	1	Overall, I am satisfied with how easy it is to use the system
	2	It was simple to use this system
	3	I could effectively complete my work quickly using this system
	4	I was able to complete my work quickly using this system
	5	I was able to efficiently complete my work using this system
	6	I felt comfortable using this system
	7	It was easy to learn to use this system
	8	I believe I could become productive quickly using this system
Information Quality (INFOQUAL)	9	The system gave error messages that clearly tell me how to fix problems
	10	Whenever I made a mistake using the system, I could recover easily and quickly
	11	The information (such as online help, on-screen messages, and other documentati on) provided with this system was clear
	12	It was easy to find the information I needed
	13	The information provided for the system was easy to understand
	14	The information was effective in helping me complete the task and scenarios
	15	The organization of information on the system screens was clear
Interface Quality (INTERQUAL)	16	The interface of this system was pleasant
	17	I liked using the interface of this system
	18	This system has all the functions and capabilities I expect it to have
Overall Usability (OVERALL)	19	Overall, I am satisfied with this system.

III. RATE OF REACTION CHEMISTRY EDUCATIONAL GAME

In this section, the basic concept of rate of reaction in high school chemistry is explained, a digital educational game for high school students focusing on the subject of chemistry is explained.

A. Rate of Reaction in Indonesian High School Chemistry

The rate of reaction in Indonesian high school chemistry concerns mainly on three main sub-topics of focus, which are collision theory, determining the order of reaction, and factors affecting the rate of reaction. Factors affecting the rate of reaction, as well as its effects are shown in Table II [38].

Based on Table II, most of the factors are easily understood due to their linear relationship. However, the enlarged surface area may be easily misunderstood due to its counterintuitive action in which grinding or smashing, i.e., a tablet, will increase its surface area. Breaking down a tablet by using a mortar and pestle into flakes, which can further be broken down into powders is multiplied for each unit of that particular object, which means a powder will have a tremendous amount of unit count compared to a single tablet, hence making it has a larger surface area. At last, an addition of catalysts affects a particular reaction when added. As such, there exist an exhaustive list of catalysts that increase the specific rate of reaction.

B. Application Design

The application design and development for this research includes 12 levels divided into four categories, the first four levels have a single factor affecting the level, and there is no special condition added, the next two levels have two different factors affecting the level, the next six levels have three different factors affecting the levels.

C. Base Gameplay and Mechanics

The common main goal of the game being designed is to improve players' understanding on factors affecting the rate of reaction in high school chemistry subject, while the main goal of the player is to be able to control the factors that affect reaction rate [22]. From a game design perspective, the goal is to create an interactive simulation game in which the players are able to control different factors that affect the rate of reaction. From an educational game design perspective, the goal focuses on knowledge acquisition of the subject of matter while also able to create a fun and engaging experience while doing so.

Fig. 1 shows the main concept of the game in which the players are able to control the factors by using an interaction, it should be able to either accelerate or decelerate the rate of reaction; hence, an indication which shows the current rate of reaction is also required.

TABLE. II. FACTORS AFFECTING THE RATE OF REACTION LEARNED IN THE HIGH SCHOOL LEVEL

Factors	Rate of reaction increases if…	The rate of reaction decreases if…
Reactant Concentration	Added	Reduced
Surface Area	Enlarged	Shrunk
Temperature	Increased	Decreased
Catalyst	Specific Catalyst Added	No Catalyst/ Incorrect Catalyst Added

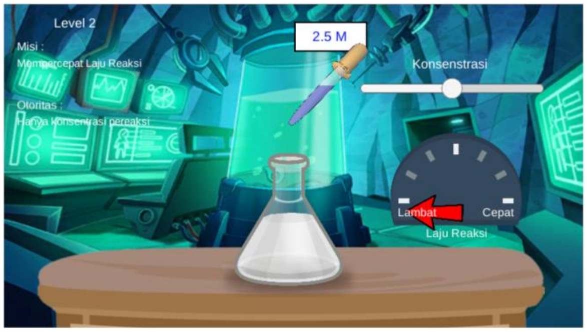


Fig. 1. Game Interface Layout.

IV. Experiment

A usability study is then executed to test the application's perceived usability. In this section, the participants of this experiment, as well as the experiment procedure, is then explained.

A. *Participants*

The participants of this study are taken from an Indonesian High School level. This study uses a total of 53 samples students. Based on its grade, the samples are divided into 2 (two) grades, 10th grade, and 11th grade, in which 10th grade has no prior knowledge regarding the topic, and 11th grade has prior knowledge regarding the topic. Based on its gender, the samples are divided into 2 (two) genders, male and female. Based on its prior material given, the samples are divided into 2 (two) groups, one with prior materials given to reinforce their study, one with no prior materials given and directly starts the game. Based on its experimental setup, the samples are divided into 2 (two) groups based on its grade, the 10th grade is done with post-test only study design, while the 11th grade is done with pretest-posttest study design. Additionally, the 10th grade has no prior knowledge regarding the rate of reaction topic while the 11th grade has prior knowledge.

B. *Procedure*

The experimental setup is done within 60-75 minutes of session divided into following:

- Ten minutes of introduction and account registration.
- Ten minutes of a pre-test quiz.
- Five minutes of re-reading subject materials (for experimental group only).
- Twenty minutes of playing the game.
- Ten minutes of a post-test quiz.
- Ten minutes of open interview and answering the usability questionnaire (CSUQ).
- Ten minutes of reserve time.

The experimental setup also sets several technical issues limitation as follows:

- Students bring and use their mobile phones.
- Students may only access the game by using Google Chrome or UC Browser.
- Students may only access the game when the session is live.

V. Results and Discussion

In this section, the reports regarding the application of CSUQ to gather questionnaire is gathered according to participants' subjective perception.

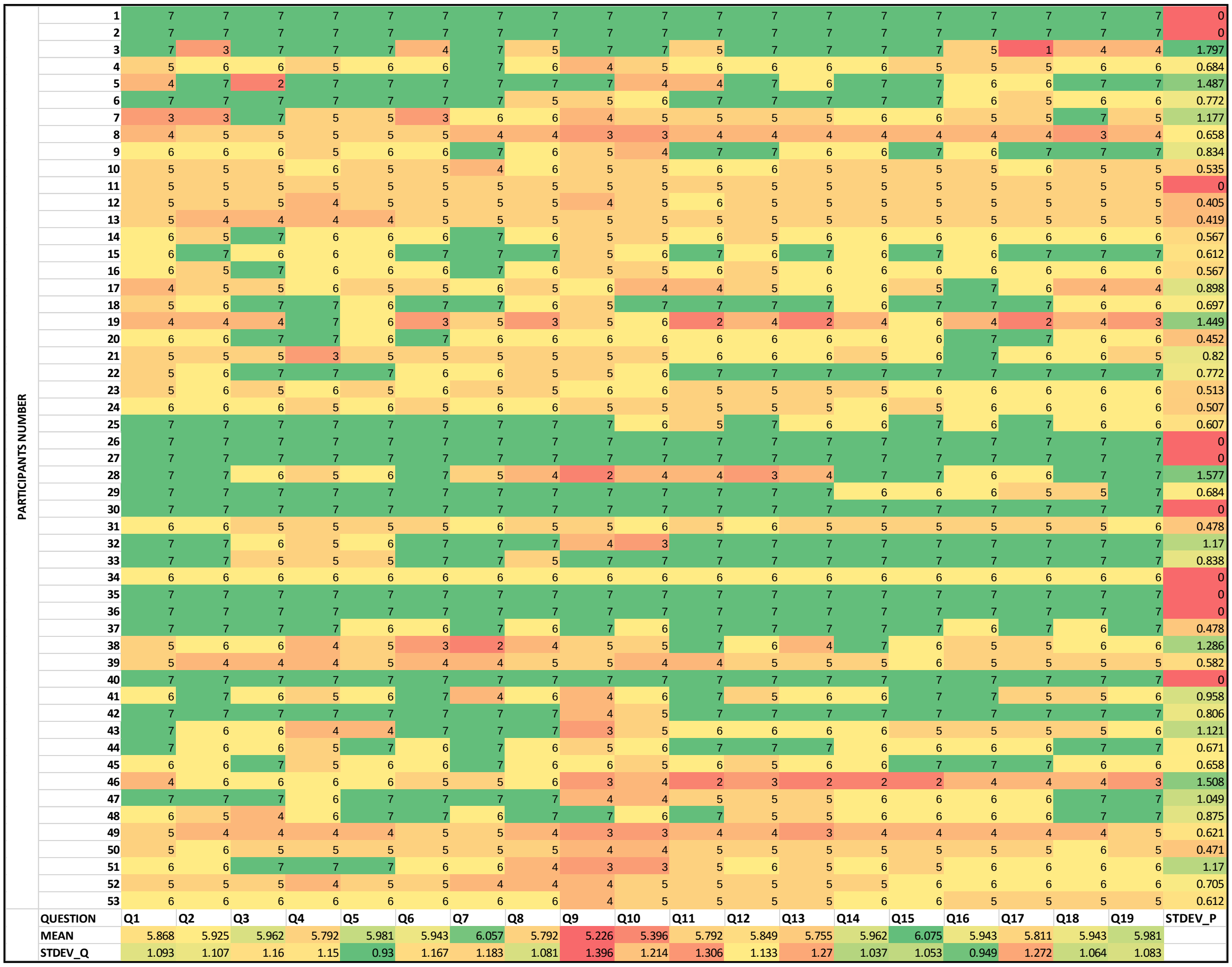
A. *General Usability*

The individual results of the usability test are shown in Fig. 2. The result indicates that Question 7 (Mean(μ) = 6.057, Standard Dev(σ) = 1.183) as well as Question 15 (Mean(μ) = 6.075, Standard Dev(σ) = 1.053) received a relatively high score compared to the other questions on the tests, suggesting that the game ease-of-use degree is relatively high as well as a relatively good organization of information, both question also has a moderate degree of standard deviation which may depict a consensus among the students.

Inversely, Question 9 (Mean(μ) = 5.226, Standard Dev(σ) = 1.396) yields a relatively low score compared to the other questions on the test, suggesting that the game wasn't able to clearly show error messages to the users, implying that the game lacks intuitiveness, although this question also shows a relatively high degree of standard deviation which may depict a much more spread-out view among the students in this test. Subsequently, Question 10 (Mean(μ) = 5.396, Standard Dev(σ) = 1.214) also indicates that the game has a degree of problem in term of error recovery, along with question 9, this shows quite a significant problem in term of how the game design displays the error and how to recover from such error. Additionally, 43 out of 53 (81.11%) participants are critical regarding the questionnaire, depicting a high degree of participation from the participants on telling their perceived usability in regards to the game.

Fig. 3 depicts the box plot of an averaged values of the CSUQ measures based on its category. Alongside that, Table III depicts the mean and standard deviation of the boxplot. In general, the result indicates a positively perceived usability from all of the participants based on the results across the categories.

System usability category (Mean(μ) = 5.915, Standard Dev(σ) = 0.858) indicates that generally, the students perceive the game to be easy to learn, simple, and useful. The information quality category (Mean(μ) = 5.722, Standard Dev(σ) = 1.002) scores relatively low compared to all of the other categories, as being stated before regarding Question 9 and 10, the lack of proper error display as well as error recovery may be one of the major issues the students are facing when using the game although the game information structure and organization is perceived to be quite good. The interface quality category (Mean(μ) = 5.899, Standard Dev(σ) = 0.999) indicates a consensus between the students that the interface is pleasant, likable, and achieved the students' expectations regarding the game. Lastly, the overall usability (Mean(μ) = 5.981, Standard Dev(σ) = 1.083) indicates that the students' general perceived usability towards the system is quite high as well.



PARTICIPANTS NUMBER	1	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	0
	2	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	0
	3	7	3	7	7	7	4	7	5	7	7	5	7	7	7	7	5	1	4	4	1.797
	4	5	6	6	5	6	6	7	6	4	5	6	6	6	6	5	5	5	6	6	0.684
	5	4	7	2	7	7	7	7	7	7	4	4	7	6	7	7	6	6	7	7	1.487
	6	7	7	7	7	7	7	7	5	5	6	7	7	7	7	7	6	5	6	6	0.772
	7	3	3	7	5	5	3	6	6	4	5	5	5	5	6	6	5	5	7	5	1.177
	8	4	5	5	5	5	5	4	4	3	3	4	4	4	4	4	4	4	3	4	0.658
	9	6	6	6	5	6	6	7	6	5	4	7	7	6	6	7	6	7	7	7	0.834
	10	5	5	5	6	5	5	4	6	5	5	6	6	5	5	5	5	6	5	5	0.535
	11	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	0
	12	5	5	5	4	5	5	5	5	4	5	6	5	5	5	5	5	5	5	5	0.405
	13	5	4	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	0.419
	14	6	5	7	6	6	6	7	6	5	5	6	5	6	6	6	6	6	6	6	0.567
	15	6	7	6	6	6	7	7	7	5	6	7	6	7	6	7	6	7	7	7	0.612
	16	6	5	7	6	6	6	7	6	5	5	6	5	6	6	6	6	6	6	6	0.567
	17	4	5	5	6	5	5	6	5	6	4	4	5	6	6	5	7	6	4	4	0.898
	18	5	6	7	7	6	7	7	6	5	7	7	7	7	6	7	7	7	6	6	0.697
	19	4	4	4	7	6	3	5	3	5	6	2	4	2	4	6	4	2	4	3	1.449
	20	6	6	7	7	6	7	6	6	6	6	6	6	6	6	6	7	7	6	6	0.452
	21	5	5	5	3	5	5	5	5	5	5	6	6	6	5	6	7	6	6	5	0.82
	22	5	6	7	7	7	6	6	5	5	6	7	7	7	7	7	7	7	7	7	0.772
	23	5	6	5	6	5	6	5	5	6	6	5	5	5	5	6	6	6	6	6	0.513
	24	6	6	6	5	6	5	6	6	5	5	5	5	5	6	5	6	6	6	6	0.507
	25	7	7	7	7	7	7	7	7	7	6	5	7	6	6	7	6	7	6	6	0.607
	26	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	0
	27	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	0
	28	7	7	6	5	6	7	5	4	2	4	4	3	4	7	7	6	6	7	7	1.577
	29	7	7	7	7	7	7	7	7	7	7	7	7	7	6	6	6	5	5	7	0.684
	30	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	0
	31	6	6	5	5	5	5	6	5	5	6	5	6	5	5	5	5	5	5	6	0.478
	32	7	7	6	5	6	7	7	7	4	3	7	7	7	7	7	7	7	7	7	1.17
	33	7	7	5	5	5	7	7	5	7	7	7	7	7	7	7	7	7	7	7	0.838
	34	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	0
	35	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	0
	36	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	0
	37	7	7	7	7	6	6	7	6	7	6	7	7	7	7	7	6	7	6	7	0.478
	38	5	6	6	4	5	3	2	4	5	5	7	6	4	7	6	5	5	6	6	1.286
	39	5	4	4	4	5	4	4	5	5	4	4	5	5	5	6	5	5	5	5	0.582
	40	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	0
	41	6	7	6	5	6	7	4	6	4	6	7	5	6	6	7	7	5	5	6	0.958
	42	7	7	7	7	7	7	7	7	4	5	7	7	7	7	7	7	7	7	7	0.806
	43	7	6	6	4	4	7	7	7	3	5	6	6	6	6	5	5	5	5	6	1.121
	44	7	6	6	5	7	6	7	6	5	6	7	7	7	6	6	6	6	7	7	0.671
	45	6	6	7	5	6	6	7	6	6	5	6	5	6	6	7	7	7	6	6	0.658
	46	4	6	6	6	6	5	5	6	3	4	2	3	2	2	2	4	4	4	3	1.508
	47	7	7	7	6	7	7	7	7	4	4	5	5	5	6	6	6	6	7	7	1.049
	48	6	5	4	6	7	7	6	7	7	6	7	5	5	6	6	6	6	7	7	0.875
	49	5	4	4	4	4	5	5	4	3	3	4	4	3	4	4	4	4	4	5	0.621
	50	5	6	5	5	5	5	5	5	4	4	5	5	5	5	5	5	5	6	5	0.471
	51	6	6	7	7	7	6	6	4	3	3	5	6	5	6	5	6	6	6	6	1.17
	52	5	5	5	4	5	5	4	4	4	5	5	5	5	5	6	6	6	6	6	0.705
	53	6	6	6	6	6	6	6	6	4	5	5	5	5	6	6	5	5	5	5	0.612
	QUESTION	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	STDEV_P
	MEAN	5.868	5.925	5.962	5.792	5.981	5.943	6.057	5.792	5.226	5.396	5.792	5.849	5.755	5.962	6.075	5.943	5.811	5.943	5.981	
	STDEV_Q	1.093	1.107	1.16	1.15	0.93	1.167	1.183	1.081	1.396	1.214	1.306	1.133	1.27	1.037	1.053	0.949	1.272	1.064	1.083	

Fig. 2. Individual Results of CSUQ Test.

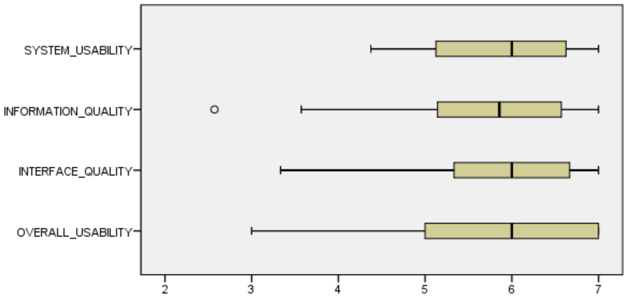


Fig. 3. Average Values of CSUQ Categories.

TABLE. III. MEAN AND STANDARD DEVIATION OF CSUQ CATEGORIES

Statistics	System Usability	Information Quality	Interface Quality	Overall Usability
MEAN	5.915	5.722	5.899	5.981
ST.DEV.	0.858	1.002	0.999	1.083

B. One-Way Analysis of Variance and Correlation

A one-way Analysis of Variance (ANOVA) has been performed to detect whether there exists any difference between the four categories of CSUQ. There was no statistically significant difference being observed as determined by the one-way ANOVA (F (4,53) = 0.6629, p= 0.5757, α=0.05) in which the p-value exceeds the α-value of 0.05. This result indicates that there is a similar perception for all the categories listed on the CSUQ test.

Table IV depicts the correlation between each CSUQ category in which the strongest correlation is observed between interface quality and overall usability (*corr*=0.845). This indicates a pleasant interface is vital for a higher overall usability score in case of an educational game system, in which the developed game has been able to reach based on previous results regarding the scores of each CSUQ category. Subsequently, a considerably strong correlation between interface quality and information quality also has been observed (*corr*=0.705). This indicates that the pleasant game interface could be reached with a proper information presentation. Although the aforementioned result is rather weak as shown in Question 9 and Question 10, a broader view of the result regarding information quality based on the average result of Question 9 to Question 15 is able to exceed and overshadow the weak result. This also indicate that no single weakness in information design in particular or game design in general that would be single-handedly responsible towards the usability score.

TABLE. IV. CORRELATION BETWEEN CSUQ CATEGORIES

	System Usability	Information Quality	Interface Quality	Overall Usability
System Usability	1			
Information Quality	0.727	1		
Interface Quality	0.681	0.705	1	
Overall Usability	0.715	0.720	0.845	1

C. Educational and Practical Insight

In addition to the CSUQ usability results, this research also gathers several educational and practical insight by synthesizing the existing CSUQ categorical results with the different grouping of students during the tests. The result can be generally split into three major categories, which are gender differences, the existence of prior materials given before gameplay, and the study design as well as prior knowledge existence regarding the topic of the game.

A general insight of this section can is presented in Fig. 4 which depicts the result of the individual questions of CSUQ questionnaire of different participants when divided into different groups based on their gender differences, the existence of prior material given, as well as the existence of prior knowledge regarding the game. Subchapters in this sections' results will explains the result in a more in-depth fashion.

In general, male and female students differs the most in term of information quality in which male students has a significantly better perceived usability whereas female students view the game information quality to be somewhat inadequate. The effect of material given before the game session affects the system usability negatively in which the group with no material given before the session perceives the game to be more useful compared to the one with material given beforehand. The group of students with no prior knowledge of the topic conveyed in the game scored much higher in their perception towards the game system usability and game simplicity, the same group also perceive the system interface quality much better compared to the group with prior knowledge, the same group also perceive the overall usability to be much higher compared to the group with prior knowledge.

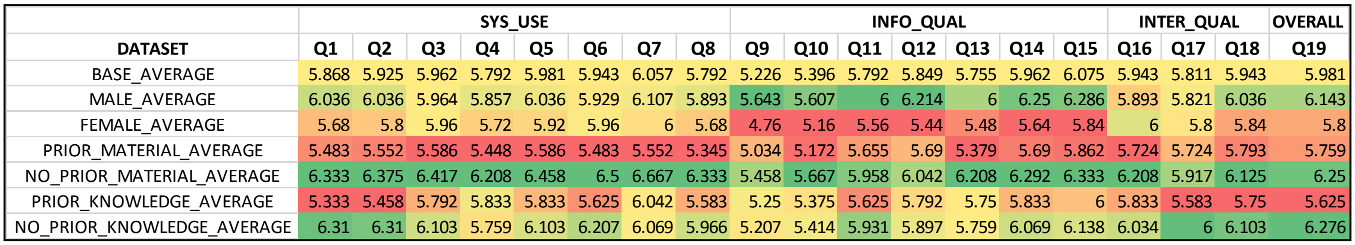
Fig. 5 depicts an interesting difference between male and female participants in this experiment. The overall usability shows that male and female students perceive the game differently. In general, male students rates the game much higher compared to their female counterparts.

As shown in the information quality category, there exist a stark contrast between male and female students scores in which the upper quartile on the male boxplot aligns with its upper whisker. This also happens on every single category in the male column. On the contrary, the female column shows more than an entire digit of difference between its median and the upper whisker. The same also happens on all other categories except the overall usability.

This result may indicate a difference between male and female students' perception towards educational game. Based on Fig. 6 and Fig. 7, male students are more likely to rate the game higher compared to their female counterpart, specifically, in general, female students rated Question 9 much lower compared to male students. This result could mean that female students require more intuitive design to cater to their expectation compared to male students. Similar result also depicted in Question 12, which male students generally rate it at least two digits higher compared to female students.

Another interesting result can also be seen in Fig. 6 on the male row which shows the median in several questions (Question 1, 2, 7, 12, 14, and 15) aligns with the upper whisker of the boxplots. This result may indicate how male students perceive educational game as a game and view it from a logical perspective by using the information given from the game much effectively compared to their female counterpart.

Based on these results, a general assumption can be made. The main difference found between male and female students are more focused on how the in-game information are being perceived as well as the perceived intuitiveness of the game. Male students may find the game to be easier to grasp compared to their female counterparts. To counteract this issue, an educational game design needs a clear depiction of information in order to improve the educational game perceived usability from female students' perspective. Hence, gender demographic in educational game may affect game design choices, especially in term of information structure in educational game.



	SYS_USE								INFO_QUAL							INTER_QUAL			OVERALL
DATASET	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19
BASE_AVERAGE	5.868	5.925	5.962	5.792	5.981	5.943	6.057	5.792	5.226	5.396	5.792	5.849	5.755	5.962	6.075	5.943	5.811	5.943	5.981
MALE_AVERAGE	6.036	6.036	5.964	5.857	6.036	5.929	6.107	5.893	5.643	5.607	6	6.214	6	6.25	6.286	5.893	5.821	6.036	6.143
FEMALE_AVERAGE	5.68	5.8	5.96	5.72	5.92	5.96	6	5.68	4.76	5.16	5.56	5.44	5.48	5.64	5.84	6	5.8	5.84	5.8
PRIOR_MATERIAL_AVERAGE	5.483	5.552	5.586	5.448	5.586	5.483	5.552	5.345	5.034	5.172	5.655	5.69	5.379	5.69	5.862	5.724	5.724	5.793	5.759
NO_PRIOR_MATERIAL_AVERAGE	6.333	6.375	6.417	6.208	6.458	6.5	6.667	6.333	5.458	5.667	5.958	6.042	6.208	6.292	6.333	6.208	5.917	6.125	6.25
PRIOR_KNOWLEDGE_AVERAGE	5.333	5.458	5.792	5.833	5.833	5.625	6.042	5.583	5.25	5.375	5.625	5.792	5.75	5.833	6	5.833	5.583	5.75	5.625
NO_PRIOR_KNOWLEDGE_AVERAGE	6.31	6.31	6.103	5.759	6.103	6.207	6.069	5.966	5.207	5.414	5.931	5.897	5.759	6.069	6.138	6.034	6	6.103	6.276

Fig. 4. Individual CSUQ Questions Result of the Participants, Separated into different Groups of Gender, Material Existence, as well Student's Prior Knowledge.

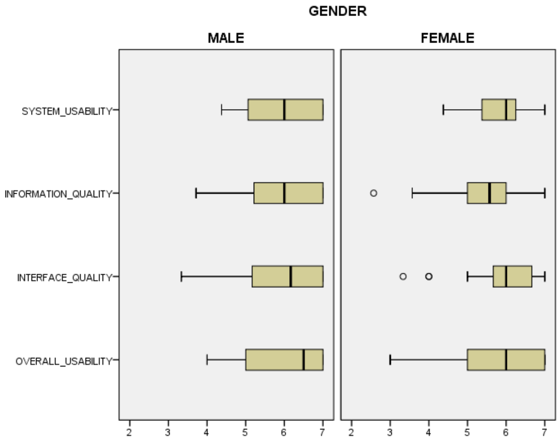


Fig. 5. Average Values of CSUQ Categories Distribution between Genders.

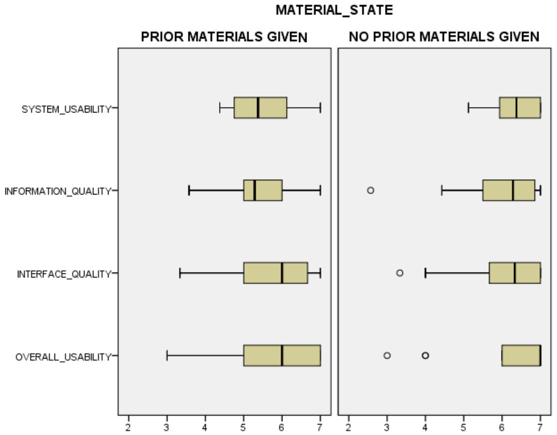


Fig. 6. Average Values of CSUQ Categories Distribution based on the Existence of Materials given before the Game Session.

The existence of materials related to the game topic given before the game session shows an interesting result. Based on Fig. 8, overall, the usability was rated lower across the board on the group with materials given beforehand while the one with no materials given beforehand has much higher rating. The difference was particularly high especially in system usability which the rating range between both groups are vastly different. This may indicate that the existence of materials given before the game session may affect the usability negatively as it may increase students' expectation towards the game. Generally speaking, sophisticated method was expected by students in this current era where smartphones is pervasive and entertainment gaming are much more graphically entertaining. The game was unable to reach such expectation and create a negative impression from the students. As mentioned before, this was particularly high in the system usability as students with materials given beforehand rated the game to be less useful as well as less sophisticated for their expectation.

In respect with Fig. 4, Fig. 9 and Fig. 10 shows that the first eight questions are rated much higher/lower depending on whether the material before the game session was given/not. Similarly, the range of scores was particularly different in Q7 in which the ease of use of the game was highly rated. The general result also shows a clear pattern that materials given before the game session affects the game usability negatively, however, a further investigation is needed whether this is just a case unique to this experiment or it is a general consensus that is adaptable to each game in existence.

Between the students' groups that has the knowledge regarding the game beforehand or not – in this case, the students have learnt about the topic of reaction rate beforehand – the usability scores show some difference in overall usability. Based on Fig. 11, the group with no prior knowledge rated the overall usability much higher compared to the ones with prior knowledge. As a student has no prior knowledge, the expectation towards the game may be lower, hence the usability rating is also higher.

Additionally, in respect to Fig. 4, Fig. 12 and Fig. 13 depicts a more specific result and difference especially in Q1 and Q2, but also in Q17 to Q19. The group with no prior knowledge perceives the game interface to be as high as their expectation as well as generally sees the game to be useful while the group with prior knowledge may either prefer the book instead or the game was simply not enticing enough for their expectation.

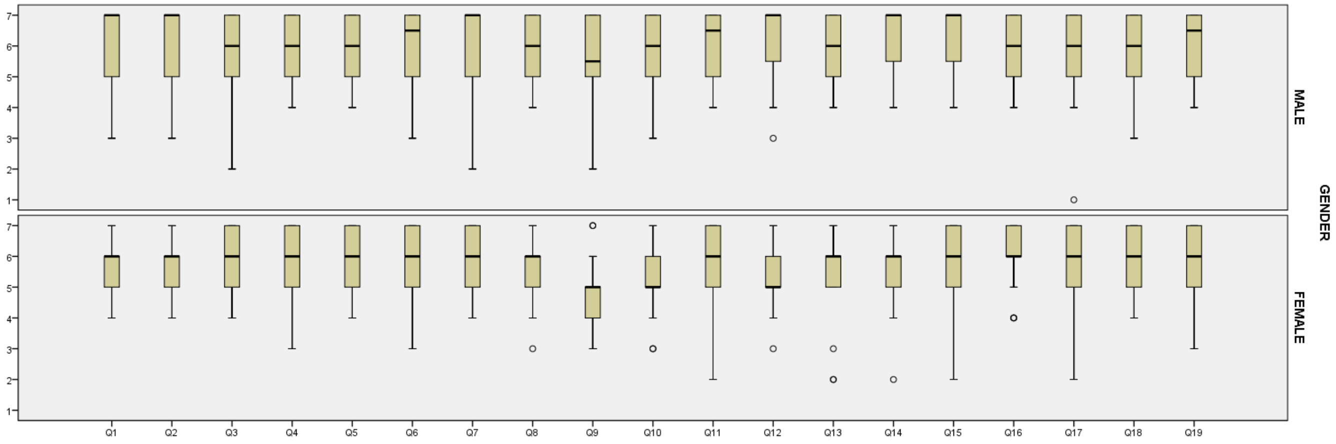
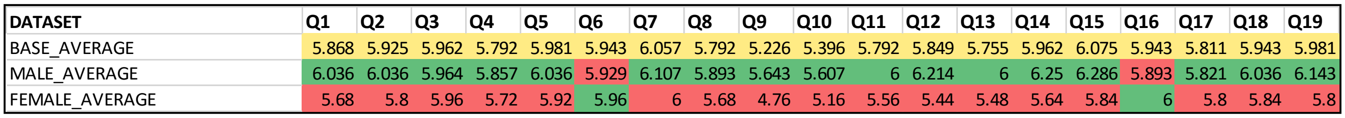


Fig. 7. Boxplot of the Average Score of CSUQ Per Questions Separated between Genders.



DATASET	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19
BASE_AVERAGE	5.868	5.925	5.962	5.792	5.981	5.943	6.057	5.792	5.226	5.396	5.792	5.849	5.755	5.962	6.075	5.943	5.811	5.943	5.981
MALE_AVERAGE	6.036	6.036	5.964	5.857	6.036	5.929	6.107	5.893	5.643	5.607	6	6.214	6	6.25	6.286	5.893	5.821	6.036	6.143
FEMALE_AVERAGE	5.68	5.8	5.96	5.72	5.92	5.96	6	5.68	4.76	5.16	5.56	5.44	5.48	5.64	5.84	6	5.8	5.84	5.8

Fig. 8. Difference in Average Score of CSUQ between Genders.

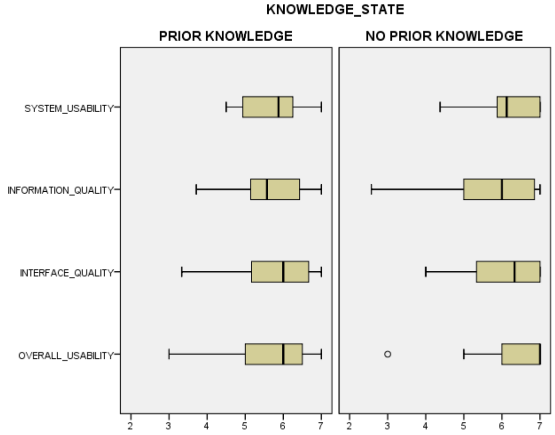


Fig. 9. Average Values of CSUQ Categories Distribution based on the Existence of Students' Prior Knowledge before the Game Session.

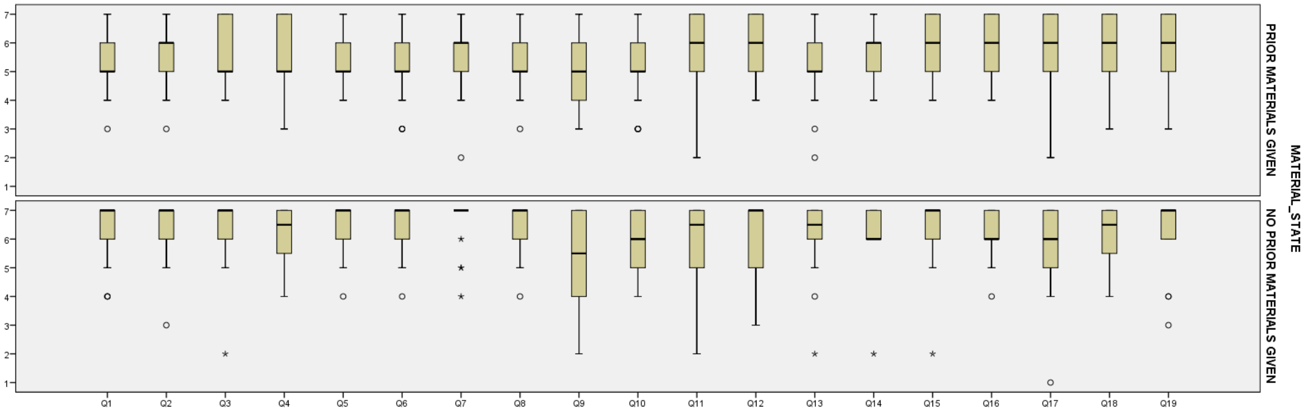
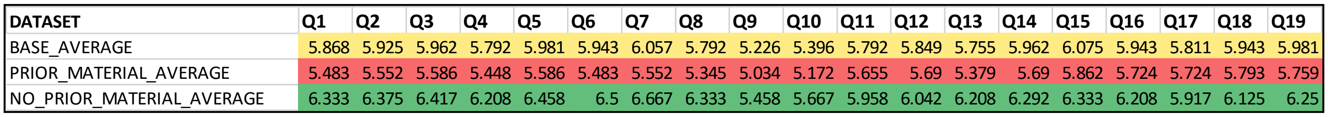


Fig. 10. Boxplot of the Average Score of CSUQ Per Questions Separated between Prior Material Existence.



DATASET	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19
BASE_AVERAGE	5.868	5.925	5.962	5.792	5.981	5.943	6.057	5.792	5.226	5.396	5.792	5.849	5.755	5.962	6.075	5.943	5.811	5.943	5.981
PRIOR_MATERIAL_AVERAGE	5.483	5.552	5.586	5.448	5.586	5.483	5.552	5.345	5.034	5.172	5.655	5.69	5.379	5.69	5.862	5.724	5.724	5.793	5.759
NO_PRIOR_MATERIAL_AVERAGE	6.333	6.375	6.417	6.208	6.458	6.5	6.667	6.333	5.458	5.667	5.958	6.042	6.208	6.292	6.333	6.208	5.917	6.125	6.25

Fig. 11. Difference in Average Score of CSUQ between Prior Material Existence.

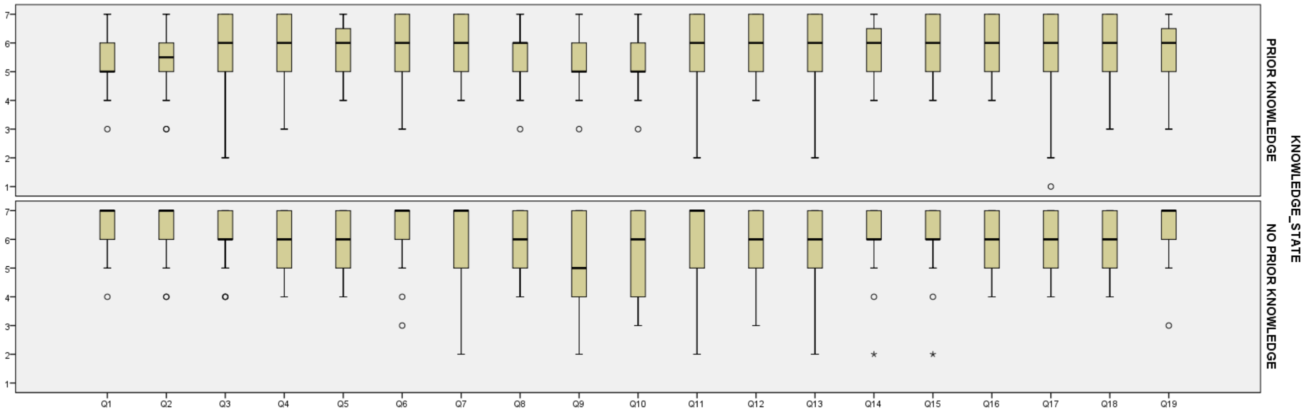
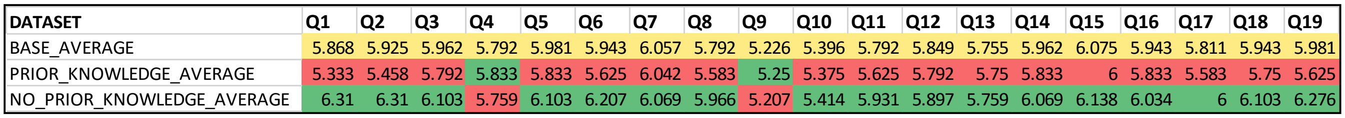


Fig. 12. More Specific Boxplot of the Average Score of CSUQ Per Questions Separated between Prior Material Existence.



DATASET	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19
BASE_AVERAGE	5.868	5.925	5.962	5.792	5.981	5.943	6.057	5.792	5.226	5.396	5.792	5.849	5.755	5.962	6.075	5.943	5.811	5.943	5.981
PRIOR_KNOWLEDGE_AVERAGE	5.333	5.458	5.792	5.833	5.833	5.625	6.042	5.583	5.25	5.375	5.625	5.792	5.75	5.833	6	5.833	5.583	5.75	5.625
NO_PRIOR_KNOWLEDGE_AVERAGE	6.31	6.31	6.103	5.759	6.103	6.207	6.069	5.966	5.207	5.414	5.931	5.897	5.759	6.069	6.138	6.034	6	6.103	6.276

Fig. 13. More Specific Difference (especially in Q1 and Q2, also in Q17 to Q19) in Average Score of CSUQ between Prior Material Existence.

VI. Conclusion and Future Works

This result presents the result of a usability testing viewed from the educational perspective of a high school chemistry educational game regarding the topic of reaction rate. The work done is contributing to the field of educational game development, specifically in terms of educational game evaluation, as well as actions that can be taken in an experimental treatment.

The initial result shows that general usability has been reached by the students as well as a high degree of correlation between each category in the usability test has been reached.

The result shows that, generally, usability scores in an educational game are affected by different gender groups, pre-game session materials, as well as prior knowledge regarding the game topic. The usability test scores show that each category in the usability tests yields a different result. In terms of system usability, the group with no materials given before the game session yields the best results. In terms of information quality, the group of male students yields the best results. In terms of interface quality, as well as overall usability, the group with no prior knowledge has a somewhat higher result.

All of the results on this, however, needs to be re-validated with a higher number of datasets to improve statistical significance, as well as done with different games to improve the validity of the result. Future works may also include more validation by relating the usability scores with students' performance as well as students' actions during the gameplay itself to see the educational effect of the game more precisely.

Acknowledgment

This research is partially funded by the Indonesian Ministry of Research, Technology, and Higher Education under WCU Program, managed by Institut Teknologi Bandung.

References

[1] J. H. Kuznekoff and S. Titsworth, "The Impact of Mobile Phone Usage on Student Learning," Commun. Educ., vol. 62, no. 3, pp. 233–252, 2013.

[2] R. Ferguson, "Learning analytics: Drivers, developments and challenges," Int. J. Technol. Enhanc. Learn., vol. 4, no. 5–6, pp. 304–317, 2012.

[3] J. L. Gómez-Urquiza, J. Gómez-Salgado, L. Albendín-García, M. Correa-Rodríguez, E. González-Jiménez, and G. A. Cañadas-De la Fuente, "The impact on nursing students' opinions and motivation of using a 'Nursing Escape Room' as a teaching game: A descriptive study," Nurse Educ. Today, vol. 72, pp. 73–76, 2019.

[4] K. Fellnhofer, "All-in-one: Impact study of an online math game for educational purposes," Int. J. Technol. Enhanc. Learn., vol. 8, no. 1, pp. 59–76, 2016.

[5] H. W. Lin and Y. L. Lin, "Digital educational game value hierarchy from a learners' perspective," Comput. Human Behav., vol. 30, no. 1, pp. 1–12, 2014.

[6] A. Fairuzabadi, A. A. Supianto, and H. Tolle, "Analysis of Players' Speed Thinking in Color Mix Game Application," Int. J. Interact. Mob. Technol., vol. 12, no. 8, pp. 113–122, 2018.

[7] A. A. Syahidi, A. A. Supianto, and H. Tolle, "Design and Implementation of Bekantan Educational Game (BEG) as a Banjar Language Learning Media," Int. J. Interact. Mob. Technol., vol. 13, no. 3, pp. 108–124, 2019.

[8] I. R. D. Renavitasari, A. A. Supianto, and H. Tolle, "Log Data Analysis of Player Behavior in Tangram Puzzle Learning Game," Int. J. Interact. Mob. Technol., vol. 12, no. 8, pp. 123–129, 2018.

[9] I. Wardani, H. Tolle, and I. Aknuranda, "Evaluation of an Educational Media on Cube Nets Based on Learning Effectiveness and Gamification Parameters," Int. J. Emerg. Technol. Learn., vol. 14, no. 14, pp. 4–18, 2019.

[10] N. M. Diah, M. Ismail, S. Ahmad, and M. K. M. Dahari, "Usability testing for educational computer game using observation method," in Proceedings - 2010 International Conference on Information Retrieval and Knowledge Management: Exploring the Invisible World, CAMP'10, 2010, pp. 157–161.

[11] O. Álvarez-Xochihua, P. J. Muñoz-Merino, M. Muñoz-Organero, C. D. Kloos, and J. A. González-Fraga, "Comparing usability, user experience and learning motivation characteristics of two educational computer games," ICEIS 2017 - Proc. 19th Int. Conf. Enterp. Inf. Syst., vol. 3, no. Iceis, pp. 143–150, 2017.

[12] J. Nielsen, Usability Engineering. Elsevier, 1994.

[13] C. Lu, M. Chang, Kinshuk, E. Huang, and C. W. Chen, "Usability of context-aware mobile educational game," Knowl. Manag. E-Learning, vol. 3, no. 3, pp. 448–477, 2011.

[14] C. G. Brown-Johnson, B. Berrean, and J. K. Cataldo, "Development and usability evaluation of the mHealth Tool for Lung Cancer (mHealth TLC): A virtual world health game for lung cancer patients," Patient Educ. Couns., vol. 98, no. 4, pp. 506–511, 2015.

[15] J. Brooke, "SUS-A quick and dirty usability scale," in Usability evaluation in industry, London, UK: Taylor & Francis, 1996, pp. 189–194.

[16] K. Finstad, "The usability metric for user experience," Interact. Comput., vol. 22, no. 5, pp. 323–327, 2010.

[17] J. R. Lewis, "IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use," Int. J. Hum. Comput. Interact., vol. 7, no. 1, pp. 57–78, 1995.

[18] F. Adnan, B. Prasetyo, and N. Nuriman, "Usability testing analysis on the Bana game as education game design references on junior high school," J. Pendidik. IPA Indones., vol. 6, no. 1, pp. 88–94, 2017.

[19] J. L. P. Medina et al., "Usability study of a web-based platform for home motor rehabilitation," IEEE Access, vol. 7, pp. 7932–7947, 2019.

[20] R. Khajouei and F. Farahani, "The evaluation of users' satisfaction with the Social Security Electronic System in Iran," Heal. Technol., pp. 1–8, 2019.

[21] S.-J. Lu, Y.-C. Liu, P.-J. Chen, and M.-R. Hsieh, "Evaluation of AR embedded physical puzzle game on students' learning achievement and motivation on elementary natural science," in Interactive Learning Environments, 2018, vol. 4820, pp. 1–13.

[22] M. Hafis, H. Tolle, A. A. Supianto, I. C. Christian, L. S. Atmojo, and S. N. Rochmainy, "Game Design Elements and Educational Game Design for Rate of Reaction Topic in High School Chemistry Subject," in 5th International Conference on Science and Technology (ICST 2019), 2019, pp. 1–6.

[23] A. A. Supianto, M. Hafis, and H. Tolle, "Significance of Dynamic Difficulty Adjustment in Delivering Instructional Scaffolding on Educational Game for Rate of Reaction Topic in High School Chemistry Subject," in 3rd International Conference on Education and Multimedia Technology (ICEMT 2019), 2019, pp. 1–6.

[24] S. Deterding, M. Sicart, L. Nacke, K. O'Hara, and D. Dixon, "Gamificaiton: Using Game Design Elements in Non-Gaming Contexts," in CHI'11 extended abstracts on human factors in computing systems, 2011, pp. 2425–2428.

[25] P. Giani and C. G. Von Wangenheim, "How to Evaluate Educational Games : a Systematic Literature Review," J. Univers. Comput. Sci., vol. 22, no. 7, pp. 992–1021, 2016.

[26] M. Prensky, "Digital Game-based Learning Prensky," Comput. Entertain., vol. 1, no. 1, pp. 1–4, 2003.

[27] M. Papastergiou, "Digital Game-Based Learning in high school Computer Science education: Impact on educational effectiveness and student motivation," Comput. Educ., vol. 52, no. 1, pp. 1–12, 2009.

[28] R. Al-Azawi, F. Al-Faliti, and M. Al-Blushi, "Educational Gamification Vs. Game Based Learning: Comparative Study," Int. J. Innov. Manag. Technol., vol. 7, no. 4, pp. 131–136, 2016.

[29] G. Bente and J. Breuer, "Why so serious? On the Relation of Serious Games and Learning," Eludamos - J. Comput. Game Cult., vol. 4, no. 1, pp. 7–24, 2010.

[30] S. Aslan and O. Balci, "GAMED: Digital educational game development methodology," Simul. Trans. Soc. Model. Simul. Int., vol. 91, no. 4, pp. 307–319, 2015.

[31] T. Stewart, "Ergonomic Requirements for Office Work with Visual Display Terminals (VDTS): Part 11: Guidance on Usability, document ISO 9241," 1998.

[32] J. R. Lewis, "Measuring Perceived Usability: The CSUQ, SUS, and UMUX," Int. J. Hum. Comput. Interact., vol. 34, no. 12, pp. 1148–1156, 2018.

[33] S. AlDakhil, E. Al Taleb, M. Al Ghamlas, and S. Al-Megren, "Assessing the Usability of a Tangible Educational Game for Children," in 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), 2019, pp. 1–5.

[34] A. Al-Sa' di, D. Parry, and P. D. Carter, "User interface preferences of young Jordanians using tablet devices," Int. J. Technol. Enhanc. Learn., vol. 10, no. 3, pp. 202–217, 2018.

[35] A. G. D. Correa, L. C. C. De Biase, E. P. Lotto, and R. D. Lopes, "Development and Usability Evaluation of an Configurable Educational Game for the Visually Impaired," 2018 IEEE Games, Entertain. Media Conf. GEM 2018, pp. 173–180, 2018.

[36] K. Maragos, "Web based Adaptive Educational Games - Exploitation in Computer Science Education," National and Kapodistrian University of Athens, 2012.

[37] M. W. M. Johnson, M. Eagle, and T. Barnes, "Invis: An interactive visualization tool for exploring interaction networks," in Educational Data Mining 2013, 2013.

[38] M. Trautz, "Das Gesetz der Reaktionsgeschwindigkeit und der Gleichgewichte in Gasen. Bestätigung der Additivität von Cv ‐ 3/2R. Neue Bestimmung der Integrationskonstanten und der Moleküldurchmesser," Zeitschrift für Anorg. und Allg. Chemie, vol. 96, no. 1, pp. 1–28, 1916.