ISSN : 2165-4069(Online)
ISSN : 2165-4050(Print)

IJARAI

International Journal of Advanced Research in Artificial Intelligence

Volume 3 Issue 10

www.ijarai.thesai.org

A Publication of
The Science and Information Organization

IJARAI

INTERNATIONAL JOURNAL OF
ADVANCED RESEARCH IN ARTIFICIAL INTELLIGENCE

SAI

THE SCIENCE AND INFORMATION ORGANIZATION

www.thesai.org | info@thesai.org

Editorial Preface

From the Desk of Managing Editor…

"The question of whether computers can think is like the question of whether submarines can swim." ― Edsger W. Dijkstra, the quote explains the power of Artificial Intelligence in computers with the changing landscape. The renaissance stimulated by the field of Artificial Intelligence is generating multiple formats and channels of creativity and innovation.

This journal is a special track on Artificial Intelligence by The Science and Information Organization and aims to be a leading forum for engineers, researchers and practitioners throughout the world.

The journal reports results achieved; proposals for new ways of looking at AI problems and include demonstrations of effectiveness. Papers describing existing technologies or algorithms integrating multiple systems are welcomed. IJARAI also invites papers on real life applications, which should describe the current scenarios, proposed solution, emphasize its novelty, and present an in-depth evaluation of the AI techniques being exploited. IJARAI focusses on quality and relevance in its publications.

In addition, IJARAI recognizes the importance of international influences on Artificial Intelligence and seeks international input in all aspects of the journal, including content, authorship of papers, readership, paper reviewers, and Editorial Board membership.

The success of authors and the journal is interdependent. While the Journal is in its initial phase, it is not only the Editor whose work is crucial to producing the journal. The editorial board members , the peer reviewers, scholars around the world who assess submissions, students, and institutions who generously give their expertise in factors small and large— their constant encouragement has helped a lot in the progress of the journal and shall help in future to earn credibility amongst all the reader members.

I add a personal thanks to the whole team that has catalysed so much, and I wish everyone who has been connected with the Journal the very best for the future.

Thank you for Sharing Wisdom!

Editor-in-Chief
IJARAI
Volume 3 Issue 10 October 2014
ISSN: 2165-4069(Online)
ISSN: 2165-4050(Print)

Editorial Board

Reviewer Board Members

CONTENTS

A Hybrid Reduction Approach for Enhancing Cancer Classification of Microarray Data

Abeer M.Mahmoud
Dept. of Computer science,
Faculty of computer and information science,
Ain Shams University, Cairo, Egypt

Basma A.Maher
Dept. of Computer science,
Faculty of computer and information science,
Ain Shams University, Cairo, Egypt

Abstract—**This paper presents a novel hybrid machine learning (ML)reduction approach to enhance cancer classification accuracy of microarray data based on two ML gene ranking techniques (T-test and Class Separability (CS)). The proposed approach is integrated with two ML classifiers; K-nearest neighbor (KNN) and support vector machine (SVM); for mining microarray gene expression profiles. Four public cancer microarray databases are used for evaluating the proposed approach and successfully accomplish the mining process. These are Lymphoma, Leukemia SRBCT, and Lung Cancer. The strategy to select genes only from the training samples and totally excluding the testing samples from the classifier building process is utilized for more accurate and validated results. Also, the computational experiments are illustrated in details and comprehensively presented with literature related results. The results showed that the proposed reduction approach reached promising results of the number of genes supplemented to the classifiers as well as the classification accuracy.**

Keywords—Mining Microarray data; Cancer classification; SVM

I. INTRODUCTION

Creatures consist of organisms and every organism carries the same genetic information. This genetic information is represented in the form of genes, where only a subset of these genes is active or expressed. Simply, Microarray gene expression data refers to such repositories of gene information that made the technology of modern biological research. Its goal is to understand the regulatory mechanism that governs protein synthesis and activity of genes. Furthermore, analyzing the gene with respect to whether and to what degree they are expressed can help characterize and understand their functions. It can further be analyzed how the activation level of genes changes under different conditions such as for specific diseases (e.g. cancers are generally caused by abnormalities in the genetic material of the transformed cells or change in their activation or function) [1].Actually, microarray represents a powerful tool in biomedical discoveries and harnessing the potential of this technology depends on the development of appropriate mining approaches [1-4].

The mining phase in the knowledge discovery process can be defined as the process of discovering interesting and unknown patterns from large amounts of data stored in information repositories [5,6]. The mining task could be one of regression, summarization, clustering and classification [5]. Classification is certainly a helpful research area in cancer diagnosis and drug discovery.

Based on the fact that microarray data is a high dimensional data with small number of samples and huge number of genes; then achieving a successful mining results with target of highly accurate and satisfied classification, the whole mining process must be divided into two main phases 1) finding the prioritized genes subset and2) building the classifier [3-6]. ML approaches achieved powerful results in data mining area. It is a bough of Artificial Intelligence (AI) that uses a variety of statistical, probabilistic and optimization methods that permit computers to “learn” from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets [5]. In the literature, there are several ML techniques for both phases. Examples of the most widely applied gene prioritized techniques for microarray data are Mean Difference (MD), Signal to noise ratio (SNR), F(x) score (FS), Fisher discriminant criterion (FC), T-test, Entropy (E), Correlation Coefficient (CC), Euclidean distance (ED), and CS [6]. Also, examples of classification techniques are Support Vector Machine (SVM), K-Nearest neighbor (KNN), Fuzzy Neural Network (FNN), and Linear Discriminate Analysis (LDA) [2-6].

In this paper, we conduct a comprehensive study that focuses on exploring and analyzing the efficiency of applying ML approaches for cancer classification. In addition, the paper mainly proposes a novel hybrid reduction approach for the enhancement of cancer classification of microarray data based on T-test, CS, KNN and SVM. The residue of this paper is primed as follows. Section 2 focuses some of related research work. Section 3 provides our methodology for reaching results of this paper. Section 4 details the four public microarray databases; with samples of their genes and their experimental settings. Computational results, comparisons & discussions are presented in section 5. Section 6 concludes the paper.

II. RELATED WORK

(2005), Wang, et al., [4]; highlighted the challenging task to choose relevant genes involved in different types of cancer. They purposed a feature selection algorithm for microarray data based on Wrappers Filters and CFS (correlation-based feature selector) and the ML algorithms such as decision trees, naïve Bayes and SVM for the classification phase. The data used in this paper was leukemia and lymphoma. F. Chu & L. Wang [7], used a SVM for cancer classification with the microarray gene expression data. The selection of genes has been completed by the use of four effective feature dimensionality reduction methods, for instance, principal components analysis (PCA), CS measure, FC, and T-test. The data set used here is SRBCT, lymphoma and leukemia data sets. The results showed that genetic selection of T-test performed well than the other three approaches. Also, in all the three data set, the SVMs obtained very good accuracies with very few numbers of genes.

(2006), Jin et al., [8]; proposed a ML techniques and used Serial Analysis of Gene Expression (SAGE) technology to facilitates and concurrently measure the expression levels of tens of thousands of genes in a inhabitants of cells. They used Chi-square is used for tag/gene selection. They investigated both binary and multi-category classification. Their experiments are performed on two human SAGE datasets: brain and breast. The results show that SVM with Chi-square is the outperforming SAGE classifier.

(2007), Wang et al., [9], proposed a new approach of two main steps. First step is gene selection, where the scoring method such as T-test, CS is used. The second one is the classification accuracy of gene combination that has been carried by using a fine classifier. Divide and conquer approach are used to attain good accuracy. Two of the datasets used in this experiment are Lymphoma Data, SRBCT Data. They used a KNN algorithm, for the treatment of missing values in microarray data. Also, they used a FNN and SVM classifier. The top marker genes are passed one by one to the classifier until good accuracy is achieved.

(2009), M. Rangasamy & S. Venketraman [10] developed a new algorithm for ranking the gene based on a classical statistical technique and two various classifiers. The paper used two types of databases, two classes datasets such as Liver and Leukemia and more than two classes database such as Lymphoma. They used a Gene selection like ANOVA, LDA and SVM-OAA RBF Kernel according to suitability of database type. Also, they used SVM-one-against-all (SVM-OAA) and LDA as a classifier for performance evaluation. The classifier is trained using all possible gene combinations; therefore the best gene combination was reported. Manuel et al., [11] presents a Kernel Alignment KNN for cancer classification using gene expression profiles. Kernel alignment KNN performs well when compared with other metric learning strategies and improves the classical KNN.

(2010), N. Revathy & R. Amalraj [12] developed a new technique that combines the enrichment score with the SVM classifier for cancer classification in microarray data. The data set is randomly divided into training and testing. The gene ranking is done then the top genes is passed into the classifier one by one if no good accuracy is attained, gene combination can be performed from the ranked data set. The performance accuracy of the SVM with the enrichment score performed well with higher accuracy than the SVM with T-Score.

(2011), Z.Ghorai et.al, [13] offered a nonparallel plane proximal classifier (NPPC) ensemble for cancer classification based on microarray gene expression profiles. A hybrid computer-aided diagnosis (CAD) framework is introduced based on filters and wrapper methods. Minimum redundancy maximum relevance (MRMR) ranking method is used for feature selection. The wrapper method is applied on those gene sets to reduce the computational burden and nonparallel plane proximal classifier (NPPC).

(2013), Abeer M. Mahmoud, et.al [14] highlighted the discovery of differentially expressed genes (DEGs) in microarray data in their way to build an accurate and cost effective classifier. A T-Test feature selection technique and KNN classifier was applied on the Lymphoma data set to reach the DEGs and to analyzing the effect of these genes on the classifier accuracy, respectively.

III. Computational Intelligence Techniques

The main objective of this paper is to successfully mine the high dimensional microarray data using ML techniques and hence propose a better approach for the mining process. The mining process will be divided into two main phases 1) finding the prioritized genes subset and2) building the classifier. Two approaches for gene ranking (T-test and CS) and two classifiers (KNN & SVM) are used. Therefore, the coming subsections presents necessary background and nomenclatures for understanding the applied ML techniques.

A. *Finding the Prioritized Genes*

Gene expression profiling or microarray analysis has enabled the measurement of thousands of genes. Identifying genes that are differentially expressed under two or more treatment conditions is a primary goal of most microarray studies. Traditionally, the methods for gene selection are broadly divided into three categories: filter, wrapper and embedded methods [15]. A filter method relies on general characteristics of the training data to select genes which show dependences on the class labels without involving any classifier for evaluation [16].They evaluate a gene based on its discriminative power for the target classes without considering its correlations with other genes. Examples are methods based on statistical ranking of individual genes, such as, correlation coefficient, t-statistics, class separability, or Fisher’s criterion, etc. [6]. The wrapper methods involve the classifiers as evaluation functions and search for the optimal gene set for classification [16]. Where training sets are used while validation set is kept separated from the training data. Therefore, the wrapper method is very slow as they search several combinations of genes and optimal parameter set and certainly adds excessive computational complexity. The embedded method performs the selection of genes during the training procedure and is specific to the particular learning algorithms [17]. This paper concatenates on the filter method, where the selected two machine learning methods for finding the differentially expressed genes are T-test & CS.

1) T-test Statistics (TS): The T-test statistics is a very famous ranking gene selection technique which is widely used by many researchers. The TS starts by calculating the Mean Difference and then normalizing it as illustrated in (1) and (2). Actually, the T-test is used to measure the difference between two Gaussian distributions. Then the P-values which define the difference significance are computed. Therefore, a threshold of P-values is used to determine a set of informative genes [6].

$$\mathrm{TS(i)} = \frac{\mu_{i1}-\mu_{i2}}{S_W\sqrt{\frac{1}{n_{s1}}+\frac{1}{n_{s2}}}} \tag{1}$$

$$S_{W^2} = \frac{(n_{s1}-1)\sigma_{i1}{}^2+(n_{s2}-1)\sigma_{i2}{}^2}{n_{s1}+n_{s2}-2} \tag{2}$$

The standard T-test is only applicable to measure the difference between two groups. Therefore, when the number of classes is more than two, we need to modify the standard T-test.

In this case, the T-test has been used to calculate the degree of difference between one specific class and the centroid of all the classes. Hence, the definition of T-test for gene i can be described from (3) to (7) [6].

$$\mathrm{TS_i} = \max\left\{\left|\frac{\bar{x}_{ik}-\bar{x}_i}{m_k s_i}\right|, k = 1,2,\dots,K\right\} \tag{3}$$

Where

$$\bar{x}_{ik} = \sum_{j\in C_k} x_{ij}/n_k \tag{4}$$

$$\bar{x}_i = \sum_{j=1}^{n} x_{ij}/n \tag{5}$$

$$s_i^2 = \frac{1}{n-K}\sum_k \sum_{j\in C_k}(x_{ij}-\bar{x}_{ik})^2 \tag{6}$$

$$m_k = \sqrt{{}^{1}/_{n_k} + {}^{1}/_{n}} \tag{7}$$

Here max {yk; k = 1; 2; . . . K} is the maximum of all yk. C_k refers to class k that includes n_k samples. x_{ij} is the expression value of gene i in sample j. $\bar{x}_{ik}$ is the mean expression value in class k for gene i. n is the total number of samples. $\bar{x}_i$ is the general mean expression value for gene i. s_i is the pooled within-class standard deviation of gene i.

2) Class Separability (CS):CS of gene i is defined as:

$$\mathrm{CS_i} = {}^{\mathrm{SB_i}}/_{\mathrm{Sw_i}} \tag{8}$$

$$\mathrm{SB_i} = \sum_{i=1}^{k}(\bar{x}_{ik}-\bar{x}_i)^2 \tag{9}$$

$$\mathrm{SW_i} = \sum_{k=1}^{k}\sum_{j\in C_k}(x_{ij}-\bar{x}_{ik})^2 \tag{10}$$

$$\bar{x}_{ik=\sum_{j\in C_k}}{}^{\bar{x}_{ij}}/_{n_k} \tag{11}$$

$$\bar{x}_i = \sum_{j=1}^{n}{}^{x_{ij}}/_{n} \tag{12}$$

Here SBi is the sum of squares of between-class distances (the distances between samples of different classes). SWi is the sum of squares of with-in class distances (the distances of samples within the same class). In the whole data set, there are K classes.C_k refers to class k that includes nk samples. $x_{ij}$is the expression value of gene i in sample j. $\bar{x}_{ij}$ is the mean expression value in class k for gene i. n is the total number of samples.$\bar{x}_i$ is the general mean expression value for gene i. CS is calculated for each gene. A larger CS indicates a larger ratio of the distances between different classes to the distances within one specific class. Therefore, CS is used to measure the capability of genes to separate different classes [9].

B. Machine Learning Classifiers

The most important application of microarray in gene expression analysis is to classify the unknown tissue samples according to their gene expression levels with the help of known sample expression levels. The small number of samples and the level of noise make the classification task of a test of challenge. In the following, two machine learning classifiers (KNN and SVM) are presented.

1) K- Nearest Neighbor (KNN): KNN is the simplest machine learning technique for classifying objects based on closest training examples in the feature space [5]. It is instance based learning. It gathers all training data and classifiers often via a majority voting, a new data point with respect to the class of its k-nearest neighbor in the given data set. KNN obtain the neighbors in the given data set. KNN obtain the neighbors for each data by using Euclidian or Mahalanobis distance between pairs of data items. Then, assign a class label to a new sample where the majority of the chosen number of neighbors belongs. Although being a simple technique, KNN shows an outstanding performance in many cases of classifying microarray gene expression. For using KNN technique three key elements are essential, (1) a set of data for training, (2) a group of labels for the training data (identifying the class of each data entry) and (3) the value of K for deciding the number of nearest neighbors [3].

2) Support Vector Machine (SVM): SVMs are widely used in many machine learning and data mining problems due to the superior performance in data analysis. The SVM algorithm is a supervised learning technique, because they exploit prior knowledge of gene to identify unknown genes. It finds the optimal hyperplane, which maximizes the minimum distance from the hyperplane to the closest training points. This feature makes SVM a powerful tool that has been used in gene expression data analysis [5]. Actually, SVM model is a representation of the samples as points in space, mapped so that the samples of the separate categories are divided by a separate area. New samples are then mapped into that same space and predicted to belong to a category based on which area they fall on. SVMs have many mathematical features that make them attractive for gene expression analysis, including their flexibility in choosing a similarity function, sparseness of solution when dealing with large data sets, the ability to handle large feature spaces, and to identify outliers [6,7].

The structure of SVM depends on kernel functions, where the most commonly used are liner and polynomial. If there are more than two classes in the data set, binary SVMs are not sufficient to solve the whole problem. To solve multi-class classification problems, the whole problem should be converted into a number of binary classification problems. Usually, there are two approaches [7]. One is the "one against all" scheme and the other is the "one against one" scheme. In "one against all", if there are N classes in the entire data set, then N independent binary classifiers are built. Each binary classifier is in charge of picking out one specific class from all the other classes. For one specific pattern, all the N classifiers are used to make a prediction. The pattern is categorized to the class that receives the strongest prediction. The prediction strength is measured by the result of the decision function.

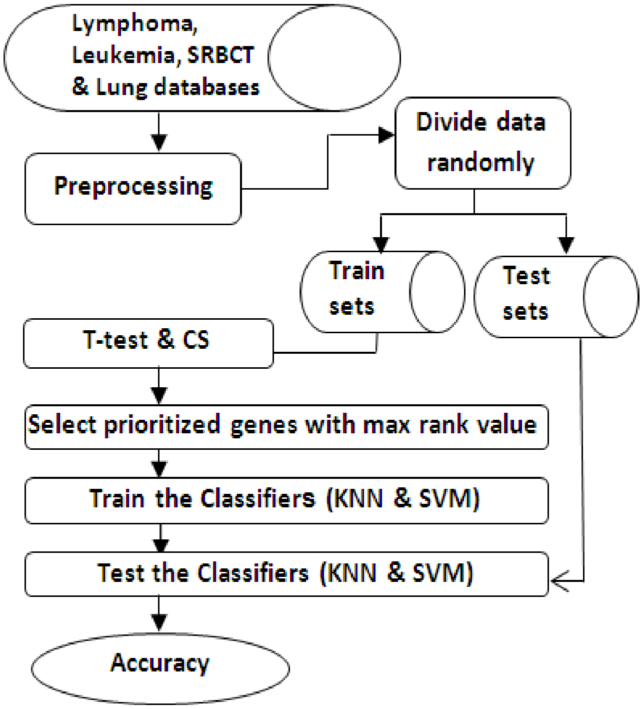

Lymphoma, Leukemia, SRBCT & Lung databases
Divide data randomly
Preprocessing
Train sets
Test sets
T-test & CS
Select prioritized genes with max rank value
Train the Classifiers (KNN & SVM)
Test the Classifiers (KNN & SVM)
Accuracy

Fig. 1. Proposed scenario for gene expression data mining workflow

For the “one against one” scheme, there must be one (and only one) classifier taking charge of the classification between any two classes. Therefore, for a data set with K classes, K(K−1)/2 binary classifiers are used. To get the ultimate result, a voting scheme is used. For every input vector, all the classifiers give their votes so there will be K(K−1)/2 votes, when all the classification (voting) finished, the vector is mapped to the class getting the highest votes. If a vector gets highest votes for more than one class, it is randomly designated to one of them [7,10]. In our practice, we choose the “one against one” scheme for database with multiclass.

C. Divcovery Workflow

Actually, to demonstrate the computational results detailed in next section, we divided the effort of this study into two main consequent lines.

In the first line,we ran all combination of applying the gene filter techniques (T-test & CS) and two ML classifiers (KNN & SVM) on four gene expression databases. Fig. 1, shows the mining workflow. First, a T-test was applied then evaluated using a KNN classifier on the three gene expression databases. Second, the prioritized genes by T-test were evaluated one more time using SVM to analyze the effect of the classifier technique on the classification accuracy. Third, CS was applied on the identical databases and evaluated by the(KNN & SVM) classifiers to analyze the effect of using different prioritized genes by different filter techniques on classification accuracy. Finally, concluding the key results.

```
Input :X ← full featured set of training data {$x_1$,..,$x_n$}
Y ← test samples
N ← Highest obtained classification accuracy  of
T-test and CS separately
            M ← control parameter of no of algorithm trails
Output : Subset$_1$ ← T-test ranked features
Subset$_2$ ← CS ranked features
            CommonSet ← Subset$_1$ ∩ Subset$_2$
T-test(X, Subset$_1$);
CS(X, Subset$_2$);
Confirm_Prioritized_Set (Subset$_1$, Subset$_2$, CommonSet)
While (C<=N) or (Count<=M)
{       Optimal(CommonSet, OptGenSet);
        Test(OptGenSet, C);
       Count ++;
}
```

Fig. 2. The Proposed hybrid reduction approach of microarray data

In the second line, as an improvement of the obtained results, we proposes a novel hybrid reduction approach for enhancing classification accuracy. A pseudo code of the proposed approach is show in Fig.2. From the figure, our proposed methodology starts by applying the T-test on the microarray data, where a ranked genes subset1 is obtained (ex: the first 100 prioritized genes). In the other hand a ranked genes subset2 is obtained from CS. A third reduction step is done by intersecting Subset1 and Subset2, where this step confirms the most important genes as a CommonSet. Then the confirmed CommonSetis searhed for the optimal genes set that enhance classification accuracy.

IV. Data Sets & Experimental Setting

For knowledge discovery in gene expression microarray data, an essential understanding of the nature of the data sets must be reached before the rest mining-workflow could proceed successfully. Hence, introductions of these databases with the experimental settings are presented in this section.

A. Microarray Genes Profile Data

Microarray datasets take the form of expression data matrix where rows represent the genes and columns represent the samples. Each cell in this data matrix is a gene expression value which expresses the gene intensity in the corresponding sample. The expression data matrix will be finally dealt with in the form Xij where; $0<i \le ng$, $0<j \le ns$ and ng, ns are the total number of genes, total number of samples respectively as in Fig. 3. [2].

Microarray data could be one of two types, paired and unpaired. Paired Data, is collected where two measurements from each patient, one before treatment and one after treatment. Then the difference between the two measurements (the log ratio) shows whether a gene has been up-regulated or down-regulated following that treatment.

$$X_{ij} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots\cdots\cdots & x1_{ns} \\ x_{21} & x_{22} & x_{23} & \vdots & x2_{ns} \\ x_{31} & x_{32} & x_{33} & \vdots & x3_{ns} \\ \cdots & \cdots & \cdots & \vdots & .. \\ \cdots & \cdots & \cdots & \vdots & .. \\ x_{ng1} & \cdots. & \cdots & \vdots & x_{ngns} \end{bmatrix}$$

Fig. 3. Expression Data Matrix

Unpaired Data, is collected where two groups of patients with two or more classes exists. To identify the genes that is up- or down-regulated in unpaired data relative to the targeted classes (i.e., differentially expressed between the two classes are selected ex: based on their statistical p-value). Therefore, the smaller p-value, the less likely it is that the observed data have occurred by chance, and the more significant the result.

B. Lymphoma dataset

The lymphoma dataset is downloaded from Lymphoma Molecular Profiling Project (LLMPP) webpage [http:// llmpp.nih.gov/lymphoma/data/figure1/figure1.cdt].This dataset contains 4026 genes and 62 samples, 42 samples derived from diffuse large B-cell lymphoma (DLBCL), 9 samples from follicular lymphoma (FL), and 11 samples from chronic lymphocytic lymphoma (CLL). The lymphoma dataset downloaded consist of noisy and inconsistent data based on its available description, ex: some additional unnecessary columns exist, and after a deep study of its important columns needed to precede our work, which are (Gene ID, Name, Class Label (DLCL, FL, CLL)), we removed such unnecessary data. Also, we found many cells values equal zero, and although we concerned reflect of such values on the classifier, but many references of our related work kept these zeros values without concern [5,6,8]. Finally, the treatment of missing attribute values (empty string), where we imputed these missing values using KNN impute technique (Matlab), where this technique replaces such data with the corresponding nearest neighbours columns and if that value is also missing, it go further to the next nearest column and so on until the treatment is achieved. Table: 1 show a sample of the Lymphoma data, where the cells in bold are the ones that were missing and then their values are imputed after pre-processing.

C. Leukemia dataset

This dataset is downloaded from the web site [http://www.broadinstitute.org/cgi-bin/cancer/publications/ pub_paper.cgi?mode%20=%20view&paper_id=43].It contains of 7129 genes and 72 samples (47 the acute lymphoblastic leukemia (ALL) samples and 25 the acute myeloid leukemia (AML) samples). The original Leukemia data was already divided into training and testing sets. There are totally 38 training samples and 34 testing samples. The 38 training samples contain 27 ALL and 11 AML. Also, the 34 testing samples contain 20 ALL and 14 AML. Actually, the downloaded leukemia dataset is already partially preprocessed where no noisy or inconsistent data exists. The available description of the leukemia dataset showed that, the only preprocessing task needed is normalization for its values to reduce the systemic bias introduced during experiments. A sample from the data is shown in Table 2.

D. The SRBCT dataset

This dataset is downloaded from [http://research.nhgri.nih .gov/microarray/Supplement/]. The SRBCT dataset is pre-divided into training and testing sets on their web site. It contains 2308 genes and 88 samples. There are totally 63 training samples and 25 testing samples. Based on its formal description, five of the testing samples doesn't belong to SRBCTs and therefore are recognized as a noisy data. These unnecessary columns are (Test 3, Test 5, Test 9, Test 11 and Test 13) [18]. The 63 training samples contain 23 Ewing families of tumors (EWS), 20 rhabdomyosarcoma (RMS), 12 neuroblastoma (NB), and 8 Burkitt lymphomas (BL). The 20 SRBCTs testing samples contain 6 EWS, 5 RMS, 6 NB, and 3 BL. A sample from the data is shown in Table 3.

TABLE I. SAMPLE OF PRE-PROCESSED LYMPHOMA DATASET

No	Gene ID	Name	Value DLCL	Value DLCL	Value DLCL	Value DLCL	Value FL	Value FL
22	GENE3069X	Clone=1340681	-1.88	**-3.3**	-2.39	**-3.3**	-0.66	**-0.66**
23	GENE2584X	Clone=1317515	-0.32	0.08	-0.08	-024	0.34	**0.34**
24	GENE3070X	Clone=1355987	-0.18	**-0.5**	-0.47	0.48	0.06	-0.14
25	GENE1843X	Clone=1268758	-0.22	0.23	**0.18**	0.52	-0.18	0.24
26	GENE3166X	Clone=1317098	-0.65	-0.26	**-0.05**	0.07	0.53	0.19
27	GENE3165X	Clone=1339226	-0.25	-0.08	**-0.32**	**0.23**	-0.12	0.1

TABLE II. A SAMPLE DATA FROM LEUKEMIA DATASET

No	Gene ID	Name	Values (ALL)	Values (ALL)	Values (AML)	Values (AML)
63	AB000114_at	Osteomodulin	72	21	39	1
64	AB000115_at	mRNA	281	250	214	103
65	AB000220_at	Semaphorin E	36	43	71	-61
66	AB000409_at	MNK1	-299	-103	-52	39
67	AB000449_at	VRK1	57	169	178	181

TABLE III. A SAMPLE DATA FORM SRBCT DATASET

No	Gene ID	Name	Values (EWS)	Values (BL)	Values (NB)	Values (RMS)
11	24145	adenylyl cyclase-associated protein	1.2607	1.4646	0.5277	0.8178
12	25584	ubiquinol-cytochrome c reductase core protein II	2.9001	2.0438	1.899	2.1544
19	29054	ARP1 homolog A	1.4482	0.8015	1.3726	1.103
20	34945	Tu translation elongation factor, mitochondrial	3.3214	1.4196	2.4937	3.0199
36	39993	superoxide dismutase1, soluble	2.1497	2.5377	1.9207	3.5434

TABLE IV. A SAMPLE DATA FROM LUNG CANCER DATASET

No	Gene ID	Values (MPM)	Values (MPM)	Values (ADCA)	Values (ADCA)
2	1000_at	214.9	249.6	60.3	202.3
3	1001_at	116.7	32.2	54	61.5
4	1002_f_at	8.4	15.2	32.6	-19.6
5	1003_s_at	-79.8	-40	-222.7	-172.4
6	1004_at	-0.3	15.3	64	18.1

TABLE V. THE PERCENTAGE OF TRAINING AND TESTING SAMPLE FOR CASE 1, CASE 2 AND CASE3

Case 1		Case 2		Case 3	
Training	**Testing**	**Training**	**testing**	**training**	**Testing**
50% (31) samples	50% (31) Samples	60% (37) samples	40% (25) samples	75% (47) samples	25% (15) samples

E. Lung Cancer Dataset

This dataset is downloaded from the web site [http://datam.i2r.astar.edu.sg/datasets/krbd/LungCancer/LungCancer-Harvard2.html]. It contains of 12533 genes and 181 samples (31 the malignant pleural mesothelioma (MPM) samples and 150 an adenocarcinoma (ADCA) samples). The original Lung Cancer data was pre-divided into training and testing. There are totally 32 training samples and 149 testing samples. The 32 training samples contain 16 MPM and 16 ADCA. Also, the 149 testing samples contain 15 MPM and 114 ADCA. A sample from the data is shown in Table 4.

V. RESULTS AND DISCUSSION

In this section, the experimental results are presented to establish the contribution of each factor used during the mining task (Phase 1: gene ranking &Phase 2: classifiers). We have conducted numerous assessments of the proposed mining workflow on four public microarray databases (Lymphoma, Leukemia, SRBCT and Lung).We have implemented it in MATLAB 7.11 (R2010b) in Windows 7 running on a PC with system configuration Intel Core 2 Duo processor (2.40 GHz) with 3 GB of RAM. Actually, to more accurately compare the performances of our applied machine learning approaches for mining microarray data, we have utilized the strategy to select genes only from the training samples. The testing samples are totally excluded from the classifier building process.

A. Phase 1: Gene Ranking and Dimensionality Reduction

1) Lymphoma dataset:We divide the lymphoma data set randomly into a three cases of subsets to study the effect of different scenarios for selecting the training and testing samples side by side with different numbers of DEGs supplemented to the classifiers on the classification accuracy. In the following, the three cases are described separately in table 5.We rank the genes by using the T-test based on their statistical score and the first 15 DEGs and their corresponding t-test values for case1, case2 and case3 is shown in table 6.

2) The SRBCT Database:Based on T-Test ranking, Table 7, shows the20 top informative 20 with their corresponding statistical values.

3) The Leukemia Database: both T-test and CS were applied o Leukemia to study the effect of different gene ranking techniques on classification accuracy. The 20 top informative genes using T-test and CS are show in table 8 with their corresponding statistical, respectively.

B. Phase 2: Classifiers

1) The SRBCT Database:it is a multiclass datase.SVM classifier deals with multiclass database in two ways, “One against one” and “one against all”, we applied “one against one” on SRBCT. The SEBCT dataset originally has four classes which are EWS, BL, NB and RMS. Therefore, applying the formula of finding the number of binary SVM classifiers K(K-1)/2, result in 6 binary classifiers. These classifiers are shown in Fig. 4 in comparison with results of applying KNN classifier on the identical testing set. Actually, during the voting scheme for obtaining the single SVM classifier, we discovered that from the six implemented and tested binary SVM classifiers of SRBCT in Fig. 4, only two classifiers (EWS & NB and BL & RMS) cover all the testing samples and hence are combined to get the average classification accuracy.

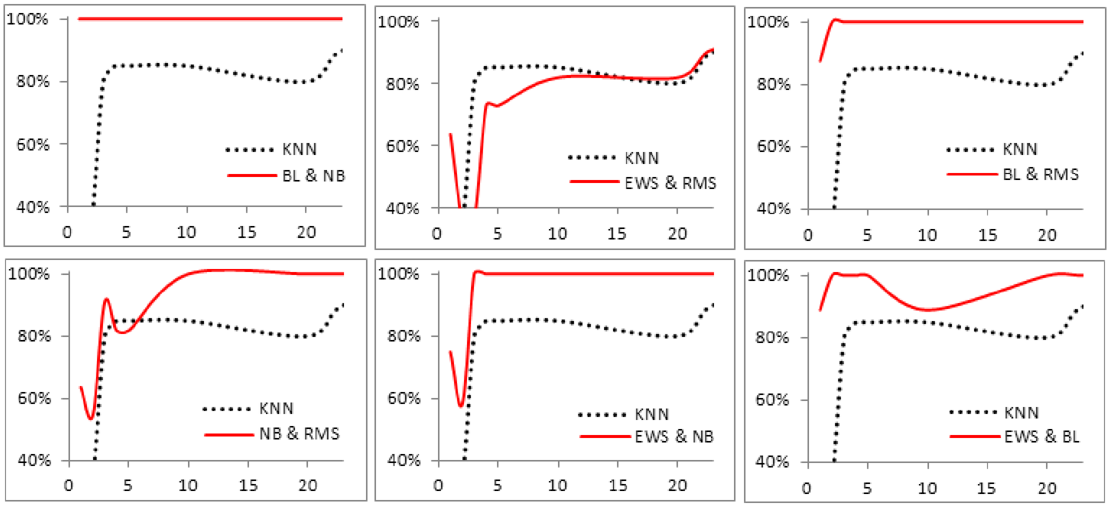


Fig. 4. six binary SVM classifiers versus KNN classifier on SRBCT database

TABLE VI. TOP 15 PRIORITIZED GENES AND THEIR CORRESPONDING T-TEST FOR CASE1,CASE2 & CASE3.

No	**Results of Case 1**		**Results of Case 2**		**Results of Case 3**	
	Gene-ID	***T-test***	***Gene-ID***	***T-test***	***Gene-ID***	***T-test***
1	Gene 639X	11.18238	Gene 653X	10.74554	Gene 653X	12.80793
2	Gene 653X	10.89338	Gene 639X	10.08582	Gene 708X	11.63526
3	Gene 769X	9.592778	Gene 563X	10.07219	Gene 699X	10.8992
4	Gene 642X	9.566121	Gene 708X	9.779857	Gene 704X	10.56475
5	Gene 2374X	8.826267	Gene 537X	9.249515	Gene 563X	10.40575
6	Gene 708X	8.775379	Gene 769X	9.091355	Gene 675X	10.33056
7	Gene 563X	8.751787	Gene 699X	8.963873	Gene 709X	10.2826
8	Gene 652X	8.667277	Gene 2203X	8.876297	Gene 706X	10.13764
9	Gene 709X	8.598637	Gene 675X	8.812909	Gene 537X	10.01828
10	Gene 704X	8.489373	Gene 704X	8.801415	Gene 639X	9.981593
11	Gene 705X	8.445308	Gene 705X	8.66119	Gene 700X	9.953322
12	Gene 2395X	8.434764	Gene 2374X	8.654025	Gene 771X	9.862814
13	Gene 2391X	8.287241	Gene 1646X	8.554017	Gene 651X	9.640458
14	Gene 721X	8.228957	Gene 700X	8.530127	Gene 2391X	9.639435
15	Gene 711X	8.193166	Gene 2395X	8.484026	Gene 540X	9.610315

TABLE VII. SRBCT 20 TOP INFORMATIVE GENES BASED ON (T-TEST)

No	**Gene ID**	**T-test Value**	**No**	**Gene ID**	**T-test Value**
1	236282	13.72563	11	745019	9.340171
2	183337	11.4937	12	609663	9.016319
3	812105	10.97844	13	325182	8.677164
4	770394	10.4634	14	784224	8.443444
5	814526	10.26562	15	68977	8.306498
6	624360	10.25416	16	769657	8.236335
7	1469292	10.02259	17	740604	8.184522
8	47475	9.939925	18	344134	8.023031
9	241412	9.790006	19	283315	7.99332
10	767183	9.471839	20	383188	7.989644

Fig. 5: shows the testing accuracy of the obtained SVM classifier in comparison with applying the KNN classifier on the same SRBCT testing samples. The figure showed that SVM recorded better classification accuracy than the KNN classifier.

2) Lymphoma dataset: For the lymphoma dataset and for each of the previously explained three cases, we applied the KNN classifier. Fig. 6: shows the testing classification accuracy for the three cases. From Fig. 6, if the highest classification accuracy is a target, then dividing training and testing subsets such case 3 is the way out, where for this dataset it reached 100% using 52 first most informative genes. But, when concerning the number of submitted genes side by side with the classification accuracy, then dividing training and testing subsets such case1 is recommended, where it reached around 85% with very few genes (less than 5 genes). For overall average classification accuracy, case 2 recorded more stability relative to changing the number of DEGs submitted to the KNN classifier.

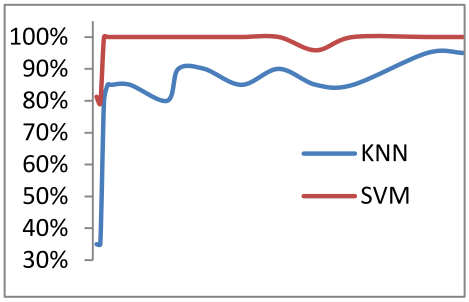


Fig. 5. Classification accuracy of KNN versus SVM classifiers on SRBCT

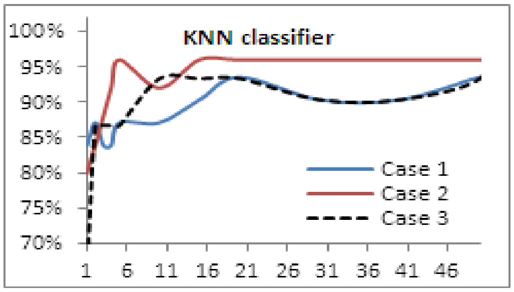


Fig. 6. KNN Testing Classification Accuracy of Cases 1,2,3 of lymphoma

TABLE VIII. LEUKEMIA 20 TOP PRIORITIZED GENES BASED ON (T-TEST & CS)

Ranked 20 genes by T-Test			Ranked 20 genes by CS	
No	**Gene-ID**	**T-test**	**Gene-ID**	**CS-value**
1	U50136_rna1_at	6.584	M55150_at	8.091
2	X95735_at	6.435	U22376_cds2_s_at	7.904
3	M55150_at	6.177	X59417_at	6.803
4	M16038_at	5.493	U50136_rna1_at	6.435
5	Y12670_at	5.488	M31211_s_at	6.293
6	M23197_at	5.387	L13278_at	6.281
7	D49950_at	5.172	U82759_at	6.268
8	X17042_at	5.042	M92287_at	6.217
9	U82759_at	5.005	U05259_rna1_at	6.181
10	M84526_at	4.952	U12471_cds1_at	6.146
11	L08246_at	4.789	U09087_s_at	6.120
12	Y00787_s_at	4.787	D26156_s_at	6.097
13	M80254_at	4.7617	X74262_at	6.016
14	U46751_at	4.7423	M81933_at	5.933
15	M27891_at	4.643	X95735_at	5.805
16	M62762_at	4.608	M28170_at	5.794
17	M63138_at	4.498	L47738_at	5.733
18	M28130_rna1_s_at	4.480	AF009426_at	5.693
19	M81695_s_at	4.414	M31523_at	5.677
20	X85116_rna1_s_at	4.338	S50223_at	5.676

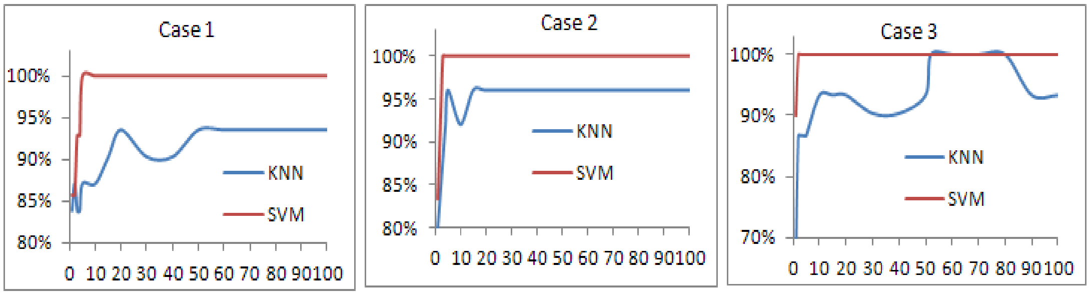


Fig. 7. Classification accuracy of KNN versus SVM on Lymphoma three cases

Actually, lymphoma is also a multiclass database with three classes (DLCL, FL and CLL), where three binary SVM classifiers (DLCL & FL, DLCL & CLL, FL &CLL)have been implemented. In a similar manner, one SVM classifier was chosen for each case of the three cases. Fig. 7: shows the classification accuracy of KNN versus SVM classifiers of every case. The figure shows that although obtaining the SVM classifier for multiclass data set took more computational effort, the SVM recorded better classification accuracy than the KNN classifier for the three cases.

3) The Leukemia Database: Fig. 8: shows the classification accuracy of KNN versus SVM classifiers for both gene ranking T-test and CS on leukemia database. The figure shows that, with few numbers of DEGs (less than 15 submitted to KNN and less than 30 genes submitted to SVM), the T-test reported better classification accuracy than CS as a ranking technique. But with greater than mentioned number of DEGs for each classifier, the CS recorded better classification accuracy.

The proposed hybrid reduction approach was applied on leukemia and resulted with CommonSet that contains 24 confirmed prioritized genes in Table 9. These common 24 genes were then searched for the optimal subset that intended to be submitted to the classifiers by applying combination and permutation. Actually, after building the first 24 SVM classifiers, where a single gene was tried at a time, the hybrid approach could reach the highest classification accuracy (94.12%) and (100%), by submitting only one gene (ranked No 6 in our CommonSet and named (M23197-at)) and integrated with SVM and KNN, respectively. Please note that (94.12%) classification accuracy was recorded before without using the proposed reduction approach and instead using the integration of (T-test+SVM) but with Subset of top ranked 30 genes instead of only one gene in our proposed reduction approach.

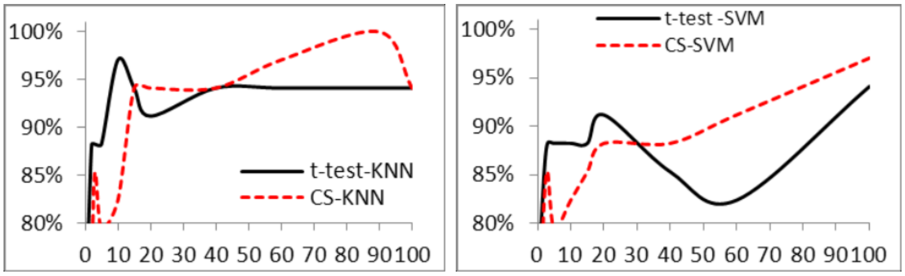


Fig. 8. T-test versus CS for both KNN & SVM classifiers on Leukemia

TABLE IX. THE24 COMMONLIST OF LEUKEMIA CANCER AND LUNG CANCER

No	Gene Name		No	Gene Name	
	Leukemia	*Lung*		*Leukemia*	*Lung Cancer*
1	U50136_rna1_at	37205_at	13	M11147_at	1030_s_at
2	X95735_at	32046_at	14	X04085_rna1_at	709_at
3	M55150_at	2047_s_at	15	M81933_at	33327_at
4	M16038_at	38482_at	16	U22376_cds2_s_at	36369_at
5	Y12670_at	37716_at	17	M86406_at	32551_at
6	M23197_at	41286_at	18	M21551_rna1_at	35822_at
7	D49950_at	40936_at	19	X15414_at	40496_at
8	X17042_at	34320_at	20	X52142_at	33328_at
9	U82759_at	33245_at	21	X59417_at	39756_g_at
10	M80254_at	39409_at	22	M31211_s_at	291_s_at
11	M62762_at	33833_at	23	D26156_s_at	34329_at
12	U12471_cds1_at	41755_at	24	L13278_at	37027_at

TABLE X. SVM & KNN CLASSIFICATION ACCURACY USING T-TEST

Data Set	Results with SVM			Results with KNN	
	(Linear SNM)	*No. of Genes*		*Accuracy (K = 1)*	*No. of Genes*
Leukemia	94.12%	6 (T-test)	70 (CS)	97.06%	10 (T-test)
				94.12%	15(CS)
Lymphoma	100% (Cases)	2 (T-test)		96% (Case2)	15 (T-test)
SRBCT	100%	3 (T-test)		90%	23 (T-test)

TABLE XI. COMPARISON OF LEUKEMIA CLASSIFIERS VERSUS GENES NUMBER

Authors	Accuracy	No. of genes
D Mishra, B Sahu, [19]	98.1%	20
Our T-tes + KNN	97.06%	10
Our T-test + SVM	94.12%	6
Our hybrid reduction approach +SVM	94.12%	1
Our hybrid reduction approach +KNN	100%	1

TABLE XII. RESULTS FOR THE SRBCT DATA SET OBTAINED BY DIFFERENT APPROACHES

Method	Accuracy	No of genes	Method	Accuracy	No of genes
MLP neural network [18]	100%	96	FNN [9]	95%	3
Nearest shrunken centroids [20]	100%	43	SVM (polynomial p=2) [7]	100%	6
Evolutionary algorithm [21]	100%	12	Our KNN	90%	23
SVM [22]	100%	20	Our SVM	100%	3

4) The Lung Cancer Database: the proposed hybrid approach reached (98.65%) with SVM using also one gene only (the ranked No 20 in the CommonSet in table 9 and named (33328_at)). Also the same classification accuracy was reached using (T-test & SVM) and without the proposed reduction approach but with subset of 16 genes. In addition, the proposed approach with KNN classifier, reached 97.31% using only one gene with rank No= 5, named (37716_at).

TABLE XIII. COMPARISON OF LYMPHOMA CLASSIFIERS

Authors	Accuracy	Number of genes
Dina A. et, al. [23]	94.59%	11
RBF SVM [7]	100%	5
Our KNN	96%	5
Our SVM	100%	2

C. Comparisons & Discussion

From table 10, 11 and Fig. 8, it is obvious that for the ***leukemia*** dataset, SVM reached 94.12% by 6 and 70 top prioritized genes by using T-test and CS ranking techniques, respectively. In addition integrating (T-test & KNN) recorded a test classification accuracy of 97.06% using 10 genes. But integrating (T-test and KNN) recorded 94.12% with gene Set=15. Therefore, based on the results (KNN+T-test) recorded higher classification accuracy. Also from table 10, the proposed approach recorded remarkable classification accuracy relative to the number of genes in comparison with almost D Mishra, B Sahu, [19].

For the ***SRBCT*** dataset, integrating (T-test& SVM), recorded 100% accuracy with 3 genes. Also, integrating (T-test& KNN) recorded 90% with 23 genes. Based on that, it is clear that SVM classifier achieved higher results on SRBCT. Actually, from table 12: and in comparison with some very related work on the same SRBCT dataset, it can be concluded that among many applied mining methods, integrating (T-test and SVM) classifier recorded best results in both classification accuracy and number of selected genes (in this case=3).

Form table 10, for the ***Lymphoma*** dataset, integrating (T-test & SVM) reached 100% in case3 with only 2 genes. Also, integrating (T-test and KNN) records 96% using 15 genes in case2. Therefore, integrating (T-test and SVM) recorded better mining results in comparison with other listed mining approaches in table 13 in terms of both classification accuracy and number of genes.

VI. Conclusion& Future Work

Biological data is known to be with a huge size; therefore mining this data is a very important research area as it deeply reflects the drug discovery, diseases diagnosis and treatment. Classifying the cancer into a predefined class based on microarray expression datasets is divided into two main phases. Phase 1 is implementing an effective gene ranking technique to reduce the number of genes involved in the classification process. Phase 2 is adjusting a powerful classifier to achieve accurate classification accuracy for new unclassified samples.

This paper presented a novel hybrid machine learning (ML) reduction approach to enhance cancer classification accuracy of microarray data based on two ML gene ranking techniques (T-test and CS). The proposed approach was integrated with two ML classifiers; KNN and SVM; for mining microarray gene expression profiles. Four public cancer microarray databases were used for evaluating the proposed approach and successfully accomplish the mining process. These were Lymphoma, Leukemia SRBCT, and Lung Cancer. The strategy to select genes only from the training samples and totally excluding the testing samples from the classifier building process was utilized for more accurate and validated results. Also, the computational experiments were illustrated in details and comprehensively presented with literature related results. Actually, integrating (T-test+SVM) recorded higher classification accuracy than the mining integrated approaches (T-test+KNN, CS+SVM, CS+KNN), where it recorded a test classification accuracy of 100% using the highest ranked 2 and 3 genes for Lymphoma and SRBCT, respectively. It also recorded 94.12% using the highest ranked 6 genes for Leukemia. The results showed that the proposed reduction approach reached promising results of the number of genes supplemented to the classifiers as well as the classification accuracy in comparison with literature similar mining approaches for microarray data.

Our future work intends to apply the proposed hybrid reduction approach on more microarray data for confirmation and verification of it performance. Also, more classifiers and ranking techniques will be studied.

Acknowledgment

The authors would like to thank Prof.Abdel Badeeh M.Salem and Prof.El-Sayed M.El-horbaty for their help, care and advices during this research.

References

[1] John N. Weinstein, et.al. The Bioinformatics of Microarray Gene Expression Profiling. Wiley-Liss, Inc,pp. 2001:46-49.

[2] Wolfgang Huber, Anja Von Hey debreck, Martin Vingron. Analysis of microarray gene expression data. Hand book of statistics genetics. 2nd edition, Wiley. 2003.

[3] Joseph S. Verducci, et.al. a Microarray analysis of gene expression: considerations in data mining and statistical treatment. Physiol. Genomics. 2006; 25(3):pp.355-363.

[4] Wang, Y., Tetko, I. -V., Hall, M. -A., Frank, E., Facius, A., Mayer, K. -F., andMewes H. -W. Gene Selection from Microarray Data for Cancer Classification—A Machine Learning Approach.ComputBiolChem.2005; 29(1):37-46.

[5] Jiawei Han, MichelineKamber. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers, An Imprint of Elsevier, First Indian Reprint. 2001.

[6] Abeer M. Mohamed, BasmaA.Maher, El-SayedM.El-horbaty& Abdel-Badeeh M. Salem. Analysis of machine learning techniques for gene selection and classification of microarray data. Proceeding of 6th IEEE int. conf. on Information Technology, Cloud Computing. 2013.

[7] F. Chu & L. Wang.Applications of Support Vector Machines To Cancer Classification With Microarray Data. International Journal of Neural Systems. 2005; 15(6):475–484.

[8] Xin Jin, A.Xu, B.Rongfang&P.Guo. Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles.Springerlink, Data Mining for Biomedical Applications. 2006; 3916:106-115.

[9] L.Wang, F.Chu, &W.Xie. Accurate Cancer Classification Using Expressions Of Very Few Genes. IEEE/ ACM Transactions on Computational Biology and Bioinformatics. 2007; 4(1):40-53.

[10] M.Rangasamy&S.Venketraman. An Efficient Statistical Model Based Classification Algorithm For Classifying Cancer Gene Expression Data With Minimal Gene Subsets. Int. J. of Cyber Society & Education.2009; 2(2):51-66.

[11] M. Martin-Merino &J.d.l.Rivas. Kernel Alignment k-NN for Human Cancer Classification Using the Gene Expression Profiles. Springer link, Artificial Neural Networks – ICANN. 2009; 5769:195-204.

[12] N. Revathy& R. Amalraj, Accurate Cancer Classification Using Expressions Of Very Few Genes.Int. J. of Computer Applications. 2010;14(4):19-22.

[13] Z.Ghorai, et.al. Cancer Classification From Gene Expression Data By NPPC Ensemble. IEEE Transactions on Computational Biology & Bioinformatics. 2011; 8(3):659-671.

[14] Abeer M. Mohamed, BasmaA.Maher, El-SayedM.El-horbaty& Abdel-Badeeh M. Salem. Applying a Statistical Technique for the Discovery of Differentially Expressed Genes in Microarray Data. Proc. Of recent advances in circuits systems, telecommunications and control, France. 2013 :220-227.

[15] Lai C, Reinders MJ, van'tVeer LJ, Wessels LF. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. BMC Bioinformatics. 2006; 7:235-253.

[16] Blanco R, Larranaga P, Inza I, Sierra B. Gene selection for cancer classification using wrapper approaches. International Journal of Pattern Recognition and Artificial Intelligence. 2004; 18(8):1373-1390.

[17] H. Peng, F. Long, and C. Ding. Feature Selection on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Trans. Pattern Analysis and MachineIntelligence. 2005; 27(8):1226-1238.

[18] J.M. Khan et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine. 2001; 7:673-679.

[19] D Mishra, B Sahu. Feature selection for cancer classification: a signal-to-noise ratio approach. International Journal of Scientific & Engineering Research. 2011; 2(4).

[20] R. Tibshirani, T. Hastie, B. Narashiman, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Nat’l Academy of Sciences USA. 2002; 99(10):6567-6572.

[21] J. Deutsch. Evolutionary algorithms for finding optimal gene sets in microarray prediction. Bioinformatics. 2003; 19(1):45-52.

[22] Y. Lee and C.K. Lee. Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics. 2003; 19(9):1132-1139.

[23] Dina A. Salem, Rania Ahmed &Hesham A. Ali. DMCA: A combined data mining technique for improving the microarray data classification accuracy. Int. Conf. on Environment and BioScience, IPCBEE. 2011; 21:36-41.

Discrimination of EEG-Based Motor Imagery Tasks by Means of a Simple Phase Information Method

Ana Loboda
Faculty of Medical Bioengineering
”Grigore T. Popa” University of Medicine and Pharmacy
Iasi, Romania

Alexandra Margineanu
Faculty of Medical Bioengineering
”Grigore T. Popa” University of Medicine and Pharmacy
Iasi, Romania

Gabriela Rotariu
Faculty of Medical Bioengineering
”Grigore T. Popa” University of Medicine and Pharmacy
Iasi, Romania

Anca Mihaela Lazar
Faculty of Medical Bioengineering
”Grigore T. Popa” University of Medicine and Pharmacy
Iasi, Romania

***Abstract*—We propose an off-line analysis method in order to discriminate between motor imagery tasks manipulated in a brain computer interface system. A measure of large-scale synchronization based on phase locking value is established. The results indicate that it can take advantage of the phase synchrony between scalp-recorded EEG activity in the supplementary motor area and in sezorimotor area, computing the differences between the active and the relaxation states. Phase locking value features are more discriminative in β rhythm than in μ rhythm. The proposed method is simple, computationally efficient and proves good results on EEG Motor Movement/Imagery Dataset available from PhysioNet research resource for physiologic signals.**

Keywords—brain computer interface; motor imagery task; electroencephalogram; phase locking value

I. INTRODUCTION

Brain computer interface (BCI) is a computerized system that acquires brain signals generated during various mental tasks, extracts and classifies suitable features, translates into appropriate control signals and sends them to an external device. BCI represents a real help for people with motor disabilities.

Various methods for monitoring brain activity (electroencephalogram - EEG, magnetoencephalogram - MEG, positron emission tomography - PET, functional magnetic resonance imaging - fMRI, single photon emission computed tomography - SPECT) can represent, by itself, the base of a brain computer interface. Taking into account the low volume and the low cost of the equipment, the simple preparation for registration, the possibility of portability, the real-time analysis, EEG is a method with certain advantages against all the others. So, it is frequently used for implementation of the BCIs.

The EEG is a non-invasive method for recording the electrical activity of the brain, using surface electrodes placed on the scalp.

According to the type of mental activities, the BCI classification is as follows:

- BCI based on evoked responses such as P300 potential, visual evoked potential;
- Operant conditioning BCI such as BCI that uses changes in cerebral rhythms, BCI using slow cortical potentials, BCI using other areas of the cortex specialized for different mental activities [1].

This research is focused on the paradigm of using as BCI discriminating tasks changes in cerebral rhythms, when a subject move or think of movement of left/right hand.

When a mental activity is produced, such as preparation, execution or imagination of hand movement, changes in the corresponding signal power of μ (8-12 Hz) and β rhythms (12-30 Hz) occur. These changes are known as event related desynchronization (ERD) and event related synchronization (ERS) of these rhythms. It was found that imagining, preparation or planning of movement develops the same kind of brain activity, ERD or ERS, in the same brain regions. When a subject intends to move or imagines to move the right or left hand, there is a short period of μ or β rhythm ERD in the opposite hemisphere of the brain. After the action period, a μ or β ERS occurs also in the opposite hemisphere of the part of the body imagined moving [2].

The main signal processing methods used in a motor imagery task paradigm are: power spectral density, parametric analysis based on autoregressive method, correlation, short Fourier transform, continuous and discrete wavelet transform [3]. Although the methods that exploit the connectivity between different regions of the brain are not so frequently used, there are some interesting papers dealing with the study of rhythms associated to motor imagery in BCI by means of phase information of the EEG signal [4], [5], [6].

We focus our attention on such a method linked with the synchronization between EEG signals from different area of the brain. Our aims are:

- to establish an algorithm suitable to decide with a high rate of success if the proposed EEG phase feature (recorded by few channels placed on the both side of the scalp) occurs when the subject has moved his left hand or when he/she moved his/her right hand;

- to see what is the rhythm that produce the better feature selection in order to discriminate between the above mentioned movement tasks.

We perform an off-line analysis in order to measure the large-scale synchronization by means of the phase locking value (PLV). Contrary to power, we expect synchrony in μ or β rhythms to be greater in the contralateral hemisphere (with respect to the target direction).

II. METHOD

There are several methods of measuring the synchronization between two signals $x(t)$ and $y(t)$. Among these methods there are worth to be mentioned phase cross-coherence analysis, mutual information, Shannon entropy, synchronization likelihood and phase locking value.

PLV is a measure of synchronization in the time domain and it is applied for analysis of EEG signals recorded during performing a motor imagery task.

PLV characterizes the stability of the phase difference between instantaneous phases $\varphi_x(t)$ and $\varphi_y(t)$ of signals $x(t)$, respectively $y(t)$ using the formula [4]:

$$PLV = \left|\langle e^{j\Delta\varphi(t)}\rangle\right| \qquad (1)$$

where $\Delta\varphi(t) = \varphi_y(t) - \varphi_x(t)$ and $\langle . \rangle$ is average operator. Usually, the averaging is done on trials, but when there is only one trial or little trials, it is performed over time (number of samples for discrete signals). In this case, PLV has the following expression:

$$PLV = \frac{1}{N}\left|\sum_{t=1}^{N} exp\left[j\left(\varphi_y(t) - \varphi_x(t)\right)\right]\right| \qquad (2)$$

where N is the number of EEG samples of the trial.

When the phase difference is constant, PLV is equal to 1. If the phase difference is distributed randomly in the interval [0, 2π], the phase difference follows a normal distribution, so PLV is equal to 0.

In order to calculate the PLV, it is necessary to know the instantaneous phases $\varphi_x(t)$ and $\varphi_y(t)$. Instantaneous phases can be obtained using wavelet transform or computing analytic signal using Hilbert transform. It is proved that there are not important differences between methods and may be considered equivalent for study of EEG signals [7]. We have chosen to use the Hilbert transform.

The Hilbert transform was defined starting from causal signals, that there is a relationship between the real and imaginary parts of Fourier transform of the signals. The Hilbert transform of a signal s(t) is given by the equation [4]:

$$\tilde{s}(t) = \frac{1}{\pi}\, p.v. \int_{-\infty}^{+\infty} \frac{s(\tau)}{t-\tau}\, d\tau, \qquad (3)$$

where p.v. is the Cauchy principal value. The analytical signal is expressed as:

$$S(t) = s(t) + j\,\tilde{s}(t) \qquad (4)$$

Instantaneous phase is calculated using the formula:

$$\varphi(t) = \arctan\left(\frac{\tilde{s}(t)}{s(t)}\right). \qquad (5)$$

It is known that there are two kind of synchrony in the brain: local-scale and large-range synchrony [8]. When dealing with adjacent channels in the same sensorimotor region of the brain it is said that local-scale synchrony may exist and when channels from widely regions are involved in computing synchrony it is said that large-scale synchrony may occur. As there are results proving best classification rates when using large-range synchronization [9], in this study we focus our attention only on this instance.

III. RESULTS

We used EEG Motor Movement/Imagery Dataset recorded using BCI2000 platform [10] available through Physionet [11]. We downloaded BCI2000 from www.bci2000.org. The database contains EEG recordings from 109 persons who performed various motor/imagery tasks. EEG recording was performed using 64 electrodes placed on the scalp according to the 10-20 extended international system (Fig. 1).

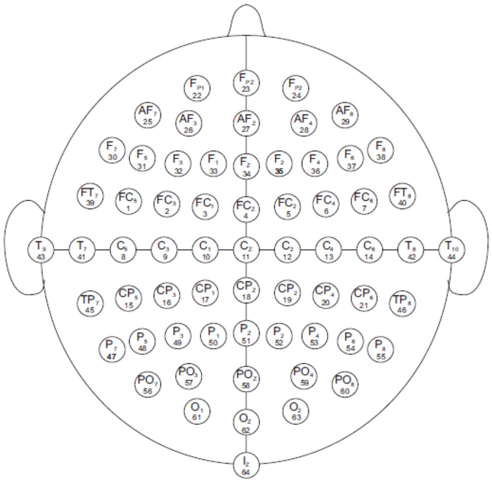


Fig. 1. Extended 10-20 international system of EEG electrodes placements

Every subject performed 14 experimental tasks: 2 runs of 1 minute for the relaxation state (one with closed eyes and one with eye open) and 3 runs of 2 minutes for each of the following tasks:

- Opening and closing the left/right fist when a target appears on the left/right side of the screen followed by relaxation.
- Imagining opening and closing the left/right fist when a target appears on the left/right side of the screen followed by relaxation.
- Opening and closing of both fists (if the target appears in the top of the screen) or of both feet (if the target appears on the lower side of the screen).
- Imagining opening and closing of both fists (if the target appears in the top of the screen) or of both feet (if the target appears on the lower side of the screen).

We used only the appropriate sets of moving and imagining movement of opening and closing the left/right fist.

The EEG recorded signals of movement and imagining of movement were grouped into three data sets: for moving denoted by 3, 7, 11 runs and for the imaginary denoted by 4, 8, 12 runs. EEG signals were sampled at a frequency of 160 Hz. We downloaded the EEG signals as *mat* files.

For each recorder there is a file with annotations of the moments when target appear in the left or the right side of the screen (indicating movement/imagining movement of right fist, left fist) and the periods of relaxation that occurs after each motor activity. Movement/imagining movement and relaxation are coded as follows: T0 relaxation period, T1 real movement/imaginary left fist, T2 real movement/imaginary right fist.

In what follows, we present the algorithm that is implemented both for real movement and for image movement of the fist. In each of these two circumstances, the study is focused on 8-12 Hz band of µ rhythm, then on 12-30 Hz of β rhythm. At the end, a comparison between results is accomplished.

In the pre-processing stage, in each of the three runs of one state, the EEG signals segmentation simultaneous to the left/right motor task or relaxation periods is performed. We used the annotations to split each run into data segments which contained the portions of the experimental run consisting of T1, T2 and T0 segments. Each segment is of 2s interval, beginning with the 0.5 s after the moment when a T1/T2 or a T0 (after T1/T2) appears in the annotation file. At this step of the script, four 3 dimension arrays (number of trials by number of samples by number of channels) are returned: one for left movement, one for right movement and two arrays for relaxation (one following the left movement and one following the right movement period). We get 7 or 8 trials for each movement task and 7 or 8 trails for the relaxation periods. There are 320 samples and 64 channels. No artifact rejection method is performed. Then, a detrending method is applied.

As one of our goal is to emphasis the effect of the frequency band upon the discrimination of the left/right motor tasks, a band-pass filter is applied in each trial. In order to avoid phase distortion, a liner phase FIR filter is used.

After the pre-processing, the representative features are extracted. The Hilbert transform is performed for all the channels and the PLV is computed for all the possible pairs of EEG channels (in all states).

From all the pairs of electrodes, we employ three electrodes from the supplementary motor area, FCz, Cz and CPz, and other six ones from the sensorimotor areas, FC3, C3, CP3 (from left hemisphere), FC4, C4 and CP4 (from right hemisphere). Therefore, we extract nine PLVs for each hemisphere, that is FCz-FC3, FCz-C3, FCz-CP3, Cz-FC3, Cz-C3, Cz-CP3, CPz-FC3, CPz-C3 and CPz-CP3 for left and FCz-FC4, FCz-C4, FCz-CP4, Cz-FC4, Cz-C4, Cz-CP4, CPz-FC4, CPz-C4 and CPz-CP4 for right hemisphere (as it is illustrated in Fig. 2).

For all these pairs, we compute the difference of PLVs between the active and relaxation periods as:

$$PLV_{diff} = PLV_{mt} - PLV_{rest} \qquad (6)$$

where $\mathrm{PLV}_{\mathrm{mt}}$ is the average PLV over all the trials in the motor task and $\mathrm{PLV}_{\mathrm{rest}}$ is the average PLV over all the trials in the relaxation state.

In order to discriminate between left or right motor activity, we compare between $\mathrm{PLV}_{\mathrm{diff}}$ of the corresponding pairs of electrodes from the left and right side and, according to the result of the comparison a vector of nine elements is built (one element for each comparison pairs). Each element of this vector may be 1 or 0 and is obtained in the following manner: if $\mathrm{PLV}_{\mathrm{diff\,left}} < \mathrm{PLV}_{\mathrm{diff\ right}}$ a value equal to 1 is put and, in the opposite case, a value equal to 0. Hereafter, a majority vote is applied for classification into the two groups of the moving/imaging left fist and the moving/imaging right fist. As it is expected the synchrony is greater in the contralateral hemisphere, the decision is as follows: if the number of elements equal to 1 in the vector is greater or equal to 5 (a half of vector length plus 1) it means that the subject moved/imagined left fist and, on the contrary, the subject moved/imagined right fist.

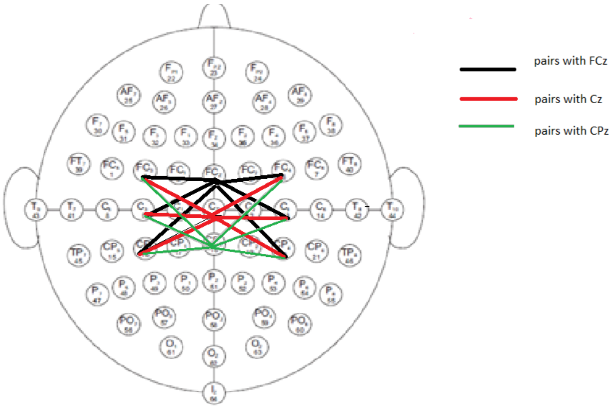


Fig. 2. The electrode pairs selection in the case of the large-range synchronization

In what follows, we denote by 100% success rate the case when, in the same run (3, 7 or 11 for movement and 4, 8 or 12 for imagery), according to our criterion, the decision that the subject moved/imagined the left/right fist corresponds for the EEG pattern extracted for the periods the target appears on the left, respectively on the right side of the screen. For example a 100% success rate is if in run 3, after the comparison of $PLV_{diff\ left}$ and $PLV_{diff\ right}$, for the movement of the target to the left side of the screen we get 6 values of 1 and for the right movement of the target we get 4 values equal to 0. Then we conclude that the left fist was moved, respectively the right fist was moved.

We worked with the mentioned database of 109 volunteers, but only for 103 of them the mentioned algorithm was performed because for the other 6 subjects there are too short or damaged records. These invalid subjects are: S043, S088, S089, S092, S100 and S104.

First of all, we focus on the implementation when the EEG signal is 8-12 Hz band-pass filtered.

The number of subjects with 100% success rate versus the number of runs is represented by bars in Fig. 3. The chart is for the two types of experiments, real and imagery movement of fist. In both circumstances, there are some subjects that have 100% success rate for all three runs (3, 7 and 11 for movement and 4, 8 and 12 for imagery), other subjects for two runs (any combinations of two runs between 3, 7 or 11 for movement or between 4, 8 and 12 for imagery) and some subjects only for one run (any of 3, 7 or 11 for movement or any of 4, 8 or12 for imagery). There are persons for whom, in the same run, the correct decision is only for the left or for the right motor task, but not simultaneously for left and right. There a cases when, in the same run, there is 100% success rate neither for the left nor for the right. For both of the last two situations, in the figure, there is mentioned by “none” on the abscise.

As we can notice from Fig. 3, there are not significant differences between the real movement and the imagery task. Besides, the number of 100% success rate for all the three runs is low, only 11/9 subjects from e total of 103 being able to correct discriminate in all runs for real movement, imagery movement, respectively. For the most of the subjects, the best result is only for two runs and one run. The worst situation, meaning that for none of the runs could be obtained 100% success rate, is for 23/21 subjects (real movement/imagery movement).

The 12- 30 Hz in another band at which we suspect phase locking to occur. So, we perform the same steps as in the previous case and chart from Fig. 4 shows the number of subjects with 100% success rate versus the number of runs.

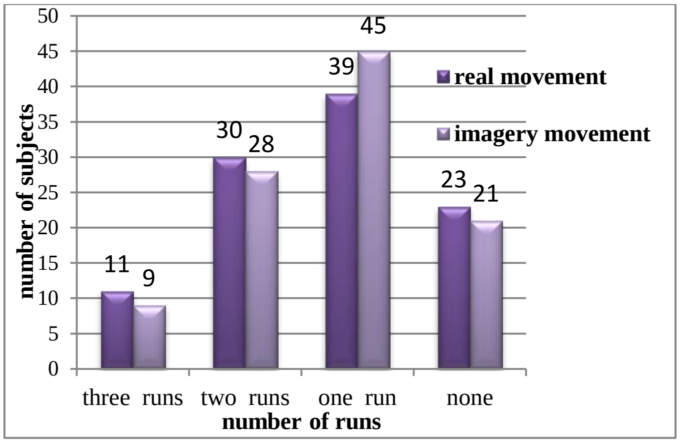


Fig. 3. Number of subjects with 100% success rate versus number of runs (EEG signal is 8-12 Hz band-pass filtered)

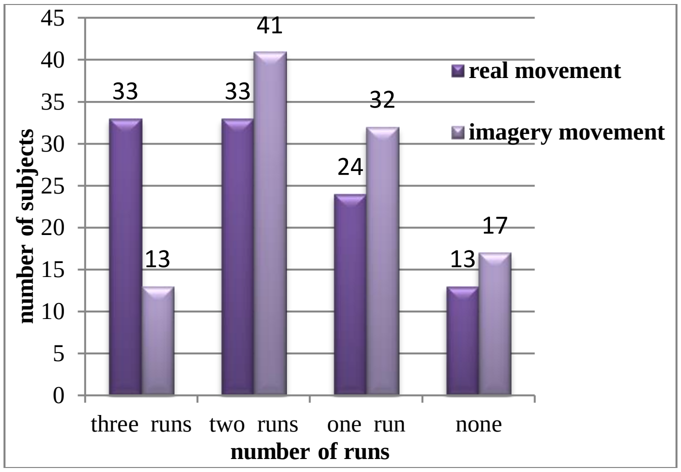


Fig. 4. Number of subjects with 100% success rate versus number of runs (EEG signal is 12-30 Hz band-pass filtered)

Comparing to Fig. 3, it is obvious that the greatest difference occurs for “three runs” in the case of real movement task, when the number of subjects who attend 100% success rate is three times higher (33 versus 11 subjects). A possible reason is that the phase locking for β rhythm is more frequent than for µ rhythm. Concerning the imagery movement task, there is no significant difference between the subjects’ number who attend 100% success rate for three runs in the case of the two different filter bands. We also observe that in the case of 12-30 Hz the number of subjects for two runs is greater than the number for one run. The number of subjects who attend 100% success rate for none of the runs is quite the same. It is possible that some of the volunteers are not able to perform the task or in the handled EEG patterns there are no discrepancies between relaxation and motor task.

It would be interesting if we compute the percentage of the left/right correct discrimination task reported to all the runs for all the subjects in the case of real movement and in imagery task. In Table 1 is depicted these results.

TABLE I. THE RATE OF CORRECT DISCRIMINATION TASK FOR ALL THE SUBJECTS

Rhythm	**Side of movement/imagery**					
	Left		***Right***		***Global***	
	Movement	***Imagery***	***Movement***	***Imagery***	***Movement***	***Imagery***
µ	64.4%	60.5%	63.1%	70.6%	63.75%	65.55%
β	75.4%	68%	82.5%	75.1%	78.95%	71.55%

In the case we filtered on 8-12 Hz (corresponding to μ rhythm), when the subjects moved the fist, the results for the two kind of tasks, left or right movement, are quite the same, taking fair values, about 64%. When the subjects imagined the movement, only for the right, the value is higher, about 70%. Better results are when the filter was on 12-30 Hz (corresponding to β rhythm), attaining 82.5% when right fist was moved.

When we compute the rate of classification in a global manner, that is considering all the matches between the position of the target (movement/imagery of the fist in that direction) and the decision of our classification method by the majority vote, we get better results for β rhythm, than for μ rhythm both for movement (78.95% versus 63.73%) and for imagery tasks (71.55% versus 65.55%).

In order to compare our method with the others, it is correct to use the same database. To our concern, the already reported researches used only the amplitude change in EEG signals for feature selection. So, in [12], despite different investigated methods (per patient/per group, normalized by frequency band/not normalized by frequency band, independent component analysis ICA/non-ICA), the reported classification accuracies do not overpass 70%, laying in 53.67%-69% interval. Taking into account these results and the fact that there were performed more difficult methods, our approach seems to be with certain advantages.

IV. Conclusions

In the framework of BCI motor imagery paradigm, because phase is supposed to include most important information about the neural activity, in order to discriminate between left and right motor task, we have proposed a phase locking value based method.

The results suggest that this method can exploit, on one hand, the phase synchrony between scalp-recorded EEG activity in the supplementary motor area and in sezorimotor area and, on the other hand, the differences between the active and the relaxation states.

PLV features were more discriminative when computed from 12-30 Hz filtered EEG signals.

The algorithm is very simple and computationally efficient using only some suitable EEG channels.

Considering the fact that no trials are excluded due to artifacts, it is a promising method for further developments of BCI systems.

V. Discussion and Future Work

In this paper we considered the average over all the trails for the PLVs for computing PLV_{diff} and, even if is time consuming, we have to test the method on each trial for every of the 103 subjects.

In order to be a valuable tool for BCI, we have to develop an appropriate method to discriminate in real time between the left and the right imagery movement.

As we used a large database, it was possible to report relevant results. There is a drawback using this database because we do not know details concerning the subjects or the experiment (e.g. the timing between runs) etc. So, we have to test the proposed method on our own EEG recordings in order to be able to explain results and establish connections to the variants that could influence them.

References

[1] A.M., Lazăr, L. Davlea, R. Ursulean, A. Maiorescu, B. Teodorescu, Interfaţa creier-calculator- Paradigme posibile, CERMI, 2009 (in Romanian)

[2] G. Pfurtscheller, C. Neuper, “Motor Imagery and direct brain-computer communication”, Proceedings IEEE, Vol. 89, No. 7, 1123-1134, 2001.

[3] A. Bashashati, M. Fatourechi, K. R. Ward and E. G., Birch “A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals”, J. Neural Eng., 4, pp. 32–57, 2007.

[4] E. Gysels and P. Celka, “Phase synchronization for the recognition of mental tasks in a brain–computer interface”, IEEE Transactions on neural systems and rehabilitation engineering, Vol. 12, No. 4, December 2004.

[5] Le Song, E. Gordon, E. Gysels, “Phase synchrony rate for the recognition of motor imagery in brain-computer interface”, Advances in Neural Information Processing Systems, vol. 18, pp. 1265-1272, 2006.

[6] D. Krusienski, D. McFarland, J. Wolpaw, “Value of amplitude, phase, and coherence features for a sensorimotor rhythm-based brain-computer interface”, Brain Res Bull., vol. 87(1), pp. 130–134, Jan 4, 2012.

[7] M. Le Van Quyen et al., “Comparison of Hilbert transform and wavelet methods for the analysis of neuronal synchrony”, Journal of Neuroscience Methods, 111, pp. 83–98, 2001.

[8] J. P. Lachaux, E. Rodriguez, J. Martinerie and F. J. Varela, “Measuring phase synchrony in brain signals”, Hum. Brain Mapp., vol. 8, pp. 194–208, 1999.

[9] Yijun Wang, Bo Hong, Xiaorong Gao, and Shangkai Gao “Phase synchrony measurement in motor cortex for classifying single-trial EEG during motor imagery”, Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA, Aug 30-Sept 3, 2006.

[10] G. Schalk, D.J. McFarland, T. Hinterberger, N. Birbaumer, J.R. Wolpaw, “BCI2000: A general-purpose brain-computer interface (BCI) system.”, IEEE Transactions on Biomedical Engineering 51(6), pp. 1034-1043, 2004.

[11] AL Goldberger et al., “PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals.” Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215]; June 13, 2000.

[12] J. Sleight, P. Pillai, and S. Mohan, "Classification of executed and imagined motor movement EEG signals," Ann Arbor: University of Michigan, pp. 1-10, 2009, Retrieved from http://www.scribd.com/doc/82045737/ICA.

Parameter optimization for intelligent phishing detection using Adaptive Neuro-Fuzzy

P. A. Barraclough
Computer Science and Digital Technology
Univeristy of Northumbria
Newcastle Upon Tyne, NE 18ST, United Kingdom

G. Sexton
Computer Science and Digital Technology
Univeristy of Northumbria
Newcastle Upon Tyne, NE 18ST, United Kingdom

M.A. Hossain
Computer Science and Digital Tecnology
University of Northumbria
Newcastle Upon Tyne, NE1 8ST, United Kingdom

N. Aslam
Computer Science and Digital Tecnology
University of Northumbria
Newcastle Upon Tyne, NE1 8ST, United Kingdom

***Abstract*— Phishing attacks has been growing rapidly in the past few years. As a result, a number of approaches have been proposed to address the problem. Despite various approaches proposed such as feature-based and blacklist-based via machine learning techniques, there is still a lack of accuracy and real-time solution. Most approaches applying machine learning techniques requires that parameters are tuned to solve a problem, but parameters are difficult to tune to a desirable output. This study presents a parameter tuning framework, using adaptive Neuron-fuzzy inference system with comprehensive data to maximize systems performance. Extensive experiment was conducted. During ten-fold cross-validation, the data is split into training and testing pairs and parameters are set according to desirable output and have achieved 98.74% accuracy. Our results demonstrated higher performance compared to other results in the field. This paper contributes new comprehensive data, novel parameter tuning method and applied a new algorithm in a new field. The implication is that adaptive neuron-fuzzy system with effective data and proper parameter tuning can enhance system performance. The outcome will provide a new knowledge in the field.**

Keywords—FIS; Intelligent phishing detection; fuzzy inference system; neuro-fuzzy

I. Introduction

Phishing is a technique utilized by attackers to obtain user's sensitive information and financial account credential for financial benefit. Phishing attacks have become a major concerned in online transactions causing monitory losses annually. According to the Press Association report, an increase in phishing attacks in online transaction caused losses of £21.6 million between January and June 2012, which was a growth of 28% from June 2011[1]. Due to this problem, various anti-phishing approaches have been proposed to solve the problem.

These approaches include feature-based techniques [2], [3], blacklist-based [4], [5], [6], [7], and content-based approaches applying machine learning algorithms have attempted to solve the problem [8], [2]. However, there is still high false positive causing inaccuracy in online transaction. The machine learning techniques also require parameter settings to solve a problem. However parameters are difficult to set to a desirable output, and parameter tuning framework are non-existent particularly for phishing website detections [9].

The main phishing website detection approaches are either utilizing: (1) Feature-based including content based approaches applying machine learning algorithms to discriminate between legitimate sites and illegitimate sites or *(*2) URL blacklist-based approach that uses a list of URL of known illegitimate websites.

This paper has made the following contributions: (1.) identified user's credential profiles as one of core component of input data that has not been utilized in the field, (2.) introduced novel data based on user's credential profile, introduced novel parameter tuning framework based on ANFIS algorithm using comprehensive feature, (3.) applied ANFIS for the first time in (phishing detection website) a new field. This is a novel work that has not been considered in literature in a unified platform.

This study focused in answering the question: how can parameter tuning method be used to maximize phishing detection accuracy using ANFIS with six sets of inputs? The aim is to design a parameter tuning method based on an adaptive neuro-fuzzy inference system, using comprehensive data from six inputs that can be used by researchers in the field. The specific objectives are: (1) to identify samples and gather comprehensive data to be used as input data, (2) to develop fuzzy models based on ANFIS comprehensive dataset, (3) to train and check/test the models using cross-validation methods, and (4) to conduct a comparative study to prove the capability and merit of the parameter tuning framework.

The outcome generated from this study should help researchers in the field with a great knowledge and understanding about the capability of fuzzy systems and six inputs.

The proposed approach applies adaptive neuro-fuzzy inference system, using six inputs including: legitimate site rules, user-behaviour profile, phishTank, user-specific sites, pop-up windows and user's credential profile. 352 data are gathered based on these six inputs. 300 data are used as input data in to the inference engine to generate fuzzy models and fuzzy rules. During 2-fold cross-validation data are split into 150 training set and 150 testing set. Trained on 150 data-set and validated on the remaining 150 set. This was repeated four-times so that data set is used only ones. This multiple experiment achieved 98.8% accuracy in real time.

Generally, phishing detections are divided into two main categories: Phishing emails and phishing websites. This study focuses on feature-based in phishing website detection, using adaptive neuro-fuzzy inference system. There are also other common machine learning algorithms that could be used including logistic regression, fuzzy logic, neural network, perceptron and many more.

The remaining sections are as follows: Section II covers literature review. Section III describes methodology including feature gathering and Analysis. Section IV covers experimental set up. Section V covers experimental set up including, training and testing. Section VI presents results and discussions and analysis. Section VII concludes the paper and provides future work.

II. RELATED WORK

Phishing attacks have increased and are becoming sophisticated, which have led to $15 billion losses in the global economy in 2012 [1]. This has caused a number of phishing solutions to be developed to tackle the problem. Anti-phishing detection solutions mainly utilize two approaches: feature-base approaches that utilize Uniform Resource Locator (URL), blacklist-based and approaches that utilize data-based including content, using machine learning techniques.

A. Content-based through Machine Learning techniques

Major researches have considered content-based approaches based on machine learning techniques to detect phishing websites [2], [10], [11], [12], [13 [14], [15]. Aburrous proposed a model to identify electronic banking sites [2]. The method utilized a combined fuzzy logic and data mining algorithms, using twenty seven characters and factors that identify phishing websites. Their approach achieved 84.4%, but suffered 15.6% error rates, which is a high risk for online users.

In an attempt to improve the detection approaches, Suriya proposed fuzzy logic, using factors and a case study to assess whether phishing attack was taking place or not [10]. Their method employed three layered checker in web pages to check for tricks of attackers, using JavaScript to hide data from users. The result revealed that their approach can detect phishing 96% correctly. However using only 3 layer method to detect phishing is limited since phishing techniques are varied.

Similarly, Wenyin considered a method based on reasoning of Semantic Link Network, using 1000 illegitimate web pages and 1000 legitimate web pages to directly discover the target name if it is a phishing website or a legitimate website [11]. Their approach had ability to identify phishing sites using inferring rules. Wenyin, however, acknowledged that the model suffered 16.6% false negative and 13.8% false positive, which are high level of error rates.

Equally, Xiang explored content-based probabilistic method that incorporates URL blacklists with shingling algorithms utilized by search engine and information retrieval technologies (IRT) to identify phishing websites [12]. Their approach had advantage of using TF-IDF and a scoring function in the search engine, when they match queries to pages that produces a probabilistic framework for detecting phishing sites. The experimental result was 67.74% and 73.53% accuracy with 0.03% error rates. Although this method has low false positives, its accuracy can make user vulnerable to phishing attacks.

Moreover, Dong focused on defending the weakest link in phishing websites detection, by analyzing online user behaviours based on visited websites and the data a user submitted to those websites [13]. Taking user's behavior into consideration is important in addressing phishing attack, but only dealing with the data users submitted to detect phishing sites is a major limitation in handling a well designed phishing websites.

Likewise, Wardman came along with a new method using file matching algorithms, hashing function index MD5 hash value and Deep MD5 Matching, to decide if a file can be utilized to classify a new file in the same group of phishing web pages [14]. Their method was tested to identify the system performance. The results demonstrated that their technique could achieve more than 90% in performance. However, the approach suffered high level of false positive rates (10%).

In the attempt to improve phishing detection scheme, Barraclough proposed a novel method to detect phishing website [15]. The approach was based on machine Neuro-fuzzy, using five sets of inputs with 288 features, which offered accuracy results of 98.4%. This result demonstrated high accuracy, but suffered 1.6% error rates. Their finding was that a hybrid neuro-fuzzy with 5 input feature-sets can detect phishing websites with high accuracy in real-time.

B. URL Blacklis-based Approaches

Another study explored blacklist-based that uses a list of URL of known illegitimate websites [4], [5], [7], [16], [17], [18], [19], [20]. For instance, Xiang proposed blacklist and content-based model to strengthen human-verified blacklist by using probabilistic techniques to obtain higher accuracy [4]. Their experiment obtained 87.42% true positive, but suffered 4.34% false positives, which is a high error rates.

Similarly, Ma conducted a study and explored phishing website detection [5]. Their approach was based on machine learning algorithms consisting of Support Vector Machine (SVM), Logistic Regression (LR) and Naïve Bayes (NB), using 10,000 host-based features from WHOIS queries with Lexical features to classify website reputation on the relationship between the lexical and host-based features. Their approach yielded 95% and 99% accuracy, and error rates range of 0.9% and 3.5%. However, Ma acknowledged that their method could not handle large evolving phishing websites that are created regularly [5].

Equally, Whittaker designed Google's phishing classifier to automate the maintenance of Google's blacklist [7]. Their method was based on logistic regression classifier, using URL-based lexical features, web page content and Hypertext Markup Language (HTML) to automatically classify phishing web pages. Their experimental results achieved 90% accuracy in real-time with 10% error rates. However, Whittaker recognized that their blacklist keeps behind with update and can only identify phishing site after it has been published and appeared on the Internet [7].

Similarly, PhishDef was developed by Le [16]. Their method was based on URLs lexical features, using algorithms to compare phishing websites. Their features were evaluated utilizing online learning algorithms including batch-based Support Vector Machine (SVM), Online Perceptron (OP), Confidence Weighted (CW) and Adaptive Regularization of Weights (AROW) that overcomes noisy data when detecting phishing websites. For each URL inputs, the classifier makes a decision whether a website is suspicious or not. Their approach achieved an average of 97% accuracy using offline algorithms and 90% using online algorithms. However, Le's research suffered features inadequacy, which is a similar problem to the study of Xiang [4]. Le's study is related to the study of Ma in their methodology. Both methods used URL feature-based [16], [5].

In addition, Huh and Kim applied search engines to measure URL which identified phishing websites and ranked them below 10, while legitimate sites were ranked top [17]. For evaluation performance, Google, Bing and Yahoo were used. As well as this, 100 legitimate websites and 100 illegitimate websites were employed, applying classification algorithms to measure website reputation including linear discrimination analysis, Naïve Bayesian, K-Nearest Neighbour and Support Vector Machine. Using K-Nearest Neighbour achieved accuracy of 95% and 6.2% error rates. Although K-Nearest Neighbour performed better in comparison with the best classifiers, URL features alone is very limited to detect phishing websites, while legitimate websites can be compromised easily by attackers and spoil their validity. Canali proposed Prophiler, a lightweight malware static filter, using HTML, JavaScript and URL with features through a classifier that identifies non-malicious pages to assess more malicious pages to a great extent [18]. While Prophiler was intended to be a fast filter, it allows higher false positive rates in order to reduce false negative rate. In addition, CANTINA+ was proposed by Xiang [19]. The approach was based on machine learning techniques, using URL, Search Engines, the HTML Document Object Model (DOM) and PhishTank with fifteen features. Although the results revealed 92% accuracy, it suffered 8% error rates. Furthermore, Ead proposed a combination of artificial immune systems and Fuzzy systems with both lexical and host-based URL features [20]. The advantage of this approach is that it classifies URLs automatically as phishing or legitimate sites.

Although the above mentioned approaches are effective to some degree of accuracy, there are still high false positive rates due to a lack of adequate data and parameter tuning methods are non-existent [9], [21]. Thus, this study address the problem: introduce a novel comprehensive data, a new parameter tuning framework and apply neuro-fuzzy system in a new field to maximize phishing detection system performance. Fig. 7 is the conceptual design of our work that illustrates the overall flow.

III. METHODOLOGY

The proposed approach consists of machine learning techniques, adaptive neuro-fuzzy and six inputs. Adaptive neuro-fuzzy is a combination of fuzzy logic and neural network. The choice of Neuro-fuzzy is that it has the advantage of both neural network which is capable of learning new data and fuzzy logic which deals with linguistic values as well as making decisions using fuzzy [If-Then] rules [9]. Six inputs include legitimate sites rules, user-behaviour profile, phishing sites, online banking sites, pop-up windows, user's credential profile. From these six sets of inputs, data are extracted that help detect phishing websites. Our phishing detection architecture with possible overall flow is presented in Fig. 1. The six inputs in part A are explained next before moving on to the fuzzy inference systems in part B.

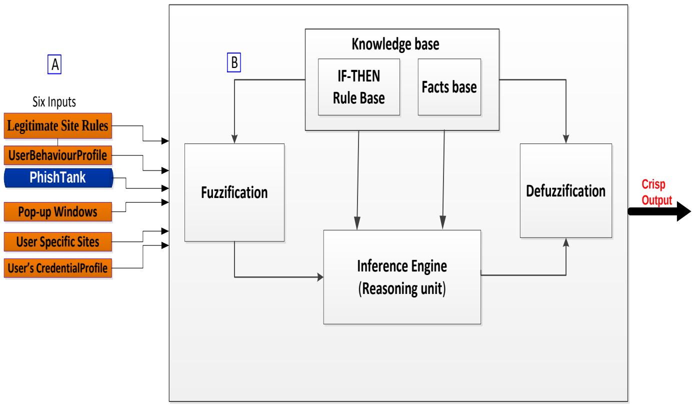


Fig.1. Fuzzy inference system for phishing website detection

A. six inputs

In part A, six inputs are diverse samples in which data are extracted, which include: Legitimate site rules, user-behaviour profile, phishTank, user-specific site, pop-up windows and user's credential profile. From these 300 data that characterize phishing techniques are gathered and are used as input data in the system to generate fuzzy models, IF-Then rules and to distinguish between phishing, suspicious and legitimates sites accurately and in real-time. The six inputs are selected carefully because they are a whole representative of phishing tactics and strategies

B. Fuzzy inference system

Part B consist of Fuzzy inference system (FIS) also called fuzzy models. Mainly, FIS for phishing detection similar to Sugeno type consist of 5 functional components: Fuzzification interface converts crisp inputs into a degree that go with linguistic value, knowledge-base is made up of rule-base that contains a number of fuzzy [IF-THEN] rules and fact-base classifies the MFs of the fuzzy sets, inference-engine performs reasoning in the decision making unit and defuzzification interface converts the fuzzy results of the inference into a crisp output.

C. Data collection and Analysis

Based on six inputs, data are randomly and carefully extracted utilizing qualitative and quantitative research method that produces numerical results. Specifically, 66 data are extracted from legitimate Site rules in the period of 23 November 2011 to 5 December 2011. A freely accessible Pinsent Manson Law Experts was consulted to identify legislations covering phishing crime and their order of importance [22]. As well as this, the European Commission documentation was explored [30]. Based on User-Behaviour profile, 60 data are extracted that cover user's information when interacting with illegitimate site (Dong et al., 2008). These data are extracted using the knowledge provided in recent journals and conference papers during the period of 8 December 2012 and 11 February 2013. Moreover, PhishTank websites provide 72 data that are extracted by exploring journal papers and 200 phishing websites from PhishTank archive [31]. Having considered that phishing techniques evolve with time, an automated wizard is utilized to extract website URLs and store in Excel Worksheets. The automated wizard also allows updates every 10 minutes when new phishing website is added into the PhishTank archive [23]. PhishTank consist of 1,038,011 verified phishing websites submitted within 3 years from 1st January 2010 to 30th December 2013. 48 data are extracted based on User-specific sites. A consultation with bank experts was done which highlighted important information and 200 legitimate bank websites were explored and compared information with phishing ones [24], [25] in the period from 8th December 2012 and 2nd February 2013. Equally, Pop-up windows consist of 42 features which are gathered by observing pop-ups that appears on websites. This was an on-going process between 28 November 2011 and 6th April 2012. 64 data are extracted from user's credential profile during the period of 8th January 2012. That makes the overall total of 352 data, also known as features in phishing terms. Data are organized in to 6 sets. In particular, set1 up to set5 which are legitimate site rules, user-behaviour, phishTank, user-specific sites and pop-up windows have been taken from our previous paper [15]. Specifically, Fig. 1 present 64 data extracted from User's credential profile that are novel and are our major contribution in this paper.

a) Data Normalization: Most frequent terms was performed across data using the 'find' function to identify data. The data is prepared using normalization method by assigning weight to each data using a value range between [0 and 1]. While 0 (zero) indicates low, 1 (one) is high and there are in between values such as 0.3. This normalization is done in order to remove deffects that occurs in data to make sure that the impact of technical bias are reduced in the results. Table I shows that data is is assigned a weight of 0.6 which indicates that the data has high importance in combating phishing, while the data weighted 0.3 is moderate, 0.1 indicates low risk.

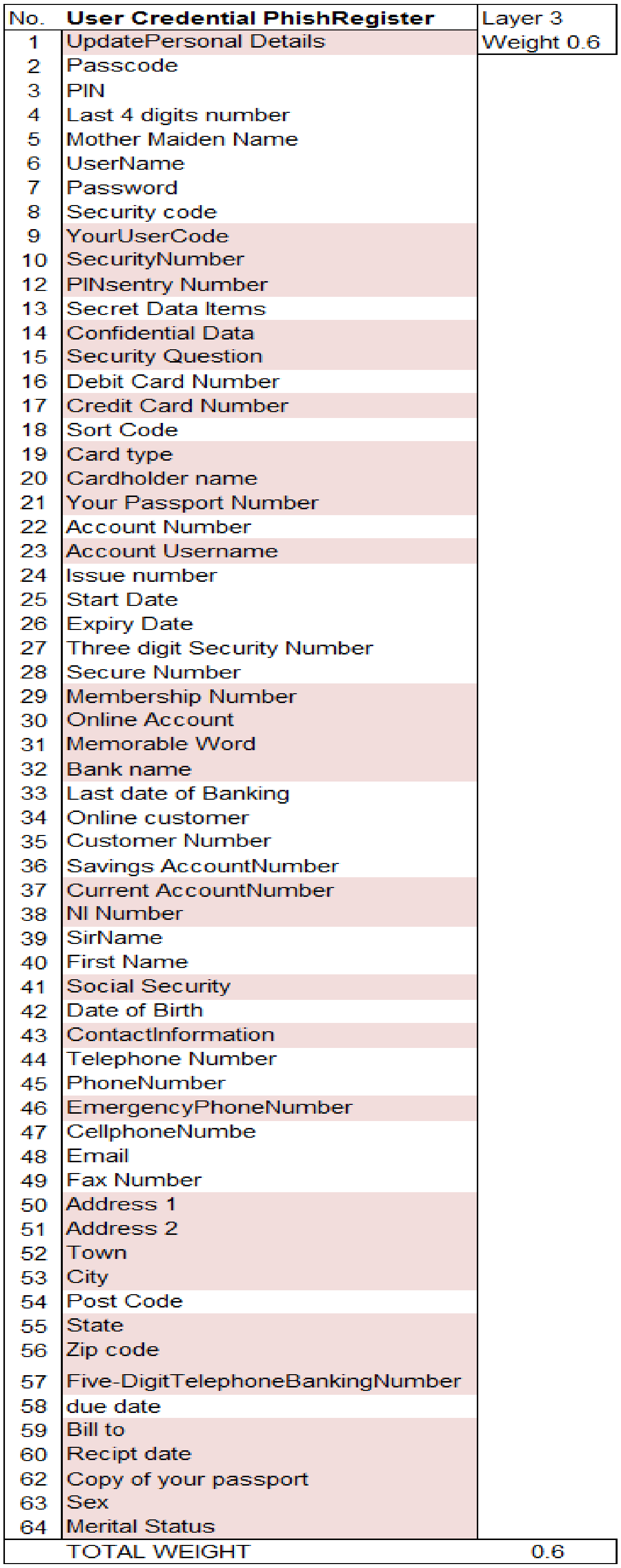
b) Feature size: Our choice of 300 data size is adequate to produce a desirable output for our model. The size of data used for modelling could be any number because the number is within the recommended range to obtain a stable cross-validation split [26]. Kohavi [27] conducted Cross-Validation experiments for accurate estimation and model selection, and found that a given number of data sets that can be partitioned into 10-fold cross-validation is good enough.

c) Methodology limitations: one of the challenge in phishing is that phishing websites are taken down within 48 hours of launching which make it hard to find them while a life. The way to solve this is to use the phishing websites published by the community service after the phishing websites have been in circulation.

IV. Experimental Set-Up

The aim of this paper is to design parameter tuning framework for phishing detection utilizing adaptive neuro-fuzzy inference system. Practically, rules are determined by expert in expert systems. In supervised learning, algorithms are trained on inputs. Thus, all input and output membership function parameters assigned are selected empirically by determining the desired input. Since there is no easy way to decide the smallest number of the hidden nodes essential to obtain a preferred level of performance, adjustments are done after evaluation if the results are not satisfying. For our experiment MATLAB fuzzy logic tool box was used because it has a FIS editor and other four integrated editors which are useful for training and testing process. Cross validation methods are used to validate the model and various Cross-validation methods exist, such as 20-Fold, 10-Fold, 5-Fold, 2-Fold and LOOCV, but 2-Fold CV is used in this paper because it can handle the conventional data well [29]. During cross-validation, 300 data is split into 150 training pair and 150 testing pair. The training pair is used to train the model, while testing pair is used for testing the model's capability. Checking, also handles the model overfitting during the training process [29].

TABLE.I. DATA EXTRACTED FROM USER'S CREDENTIAL PROFILE



No.	**User Credential PhishRegister**	Layer 3
1	UpdatePersonal Details	Weight 0.6
2	Passcode	
3	PIN	
4	Last 4 digits number	
5	Mother Maiden Name	
6	UserName	
7	Password	
8	Security code	
9	YourUserCode	
10	SecurityNumber	
12	PINsentry Number	
13	Secret Data Items	
14	Confidential Data	
15	Security Question	
16	Debit Card Number	
17	Credit Card Number	
18	Sort Code	
19	Card type	
20	Cardholder name	
21	Your Passport Number	
22	Account Number	
23	Account Username	
24	Issue number	
25	Start Date	
26	Expiry Date	
27	Three digit Security Number	
28	Secure Number	
29	Membership Number	
30	Online Account	
31	Memorable Word	
32	Bank name	
33	Last date of Banking	
34	Online customer	
35	Customer Number	
36	Savings AccountNumber	
37	Current AccountNumber	
38	NI Number	
39	SirName	
40	First Name	
41	Social Security	
42	Date of Birth	
43	ContactInformation	
44	Telephone Number	
45	PhoneNumber	
46	EmergencyPhoneNumber	
47	CellphoneNumbe	
48	Email	
49	Fax Number	
50	Address 1	
51	Address 2	
52	Town	
53	City	
54	Post Code	
55	State	
56	Zip code	
57	Five-DigitTelephoneBankingNumber	
58	due date	
59	Bill to	
60	Recipt date	
62	Copy of your passport	
63	Sex	
64	Merital Status	
	TOTAL WEIGHT	0.6

A. *Parameter Framework Descriptions*

Parameter tuning framework for intelligent phishing detection is presented in Table II. It shows parameter optimal specification that has impact in fuzzy system performances. The parameters are assigned as follows: Membership Function is assigned 4 values in column 2. Input membership function (MF) is assigned Gbell shape in column 3. Column 4 demonstrates that output membership functions are linear. 16 epochs are assigned as shown in column 5 which presents the number of iterations. The number of tolerance is assigned to 0.01 in column 6. 150 training set are assigned in column 8, while 150 validation sets are assigned as shown in column 9. The experiment is run multi-times using two-fold cross-validation method as illustrated in column 10. This process is summarized in the next section. The results and analysis of this experiment are presented in section 5 and 6 and the best performance is also highlighted.

B. *Parameter Framework Summaries*

- Step 1: A total of 300 data are utilized in Framework, which are split into 150 training set and 150 test set. The training set is utilized to generate a model and to train the fuzzy model while the remaining 150 set is utilized for testing the model.
- Step 2: 4 membership functions values are assigned for the input.
- Step 3: Linear is set for the output membership functions.
- Step 4: Parameter optimization methods are assigned to hybrid, back-propagation and least square
- Step 5:16 epochs are assigned so that after 0.01 iterations, the process stops at the minimal error tolerance which is assigned to zero tolerance.

C. *Training*

To perform training and testing for the parameter tuning framework, Cross validation (CV) methods as mentioned above is applied to train and test the parameter tuning framework models. Using 2-Fold CV, data is randomly split into training and testing sets. 2-Fold cross-validation method is used since it can handle conventional data well given the 300 data-set [29]. While training set is used to train the model. Testing set is used to check the generalization and capability of the fuzzy models and to handle over-fitting that occur during training process.

D. *Adaptive neuro-fuzzy Inference Architecture for Phishing Detection*

A model similar to Sugeno type is generated and presented in Fig. 2. The structure consists of five functional components: Input Layer, Fuzzification, Rule base, Normalisation, and defuzzification [9]. ANFIS is a multilayer neural network and applies conventional learning algorithms including back-propagation when training set is present. The processes of learning and fuzzy reasoning performed by ANFIS based on rules include:

a) Layer 1: This is the input layer. Neuron in this step simply transmits crisp straight to the next layer.

b) Layer 2: is fuzzification. In this layer, inputs are taken and classified into a degree of membership functions in which they belong as fuzzy sets. This is shown in Fig. 3.

c) Layer 3: is a Rule base where all the rules are assigned weight between [0 and 1]. For every rule, implication is implemented that generates qualified consequent as a fuzzy set of each rule depending on the firing strength. A rules-base sample containing 5 fuzzy IF-THEN rules generated through experiments is presented in Fig. 4.

d) Layer 4: is Aggregation. In this layer, each rule is combined to make a decision. The output of the aggregation process is a fuzzy set whose membership function assigns a weighting for each output value.

e) Layer 5: is defuzzification. In this layer, the input for the defuzzification process is acombined output fuzzy set and the output is a single number. The most common defuzzify method is the centroid calculation [9].

Fig. 2, a fuzzy model shows that given the values of premise parameters, the overall output is expressed as *linear* combining consequent parameters. Hybrid learning algorithm is used as parameter optimization method to enhance performance. In the forward pass for that particular algorithm, functional signals move forward until layer 4. Then consequent parameters are classified by the least square estimate (LSE). The error rates in the backward pass get propagated backward, while the premise parameters get updated using the gradient descent [9].

TABLE.II. PARAMETERS FRAMEWORK ASSIGNED

							300 Data Set		**Cross-Validation**
Parameters & Dataset	**MFs value**	**Input MFs**	**Output MFs**	**Para. Optimization**	**No. of Epoch**	**No. of Tolerance**	**Training set**	**validation set**	2-Fold
Assigned Value	4	Gbell	Linear	Hybride	16	0.01	150	150	

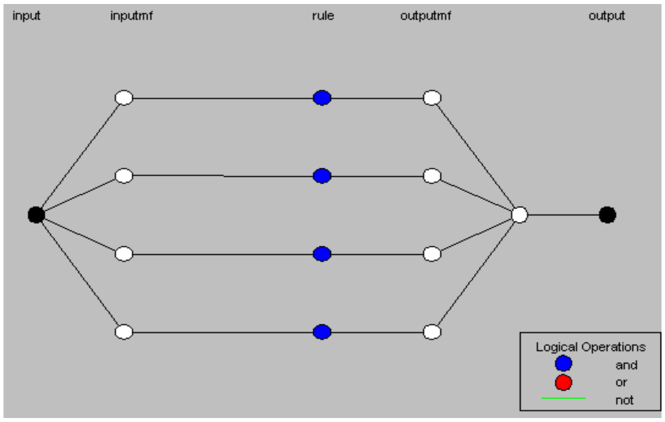


Fig.2. Fuzzy inference model for detecting phishing

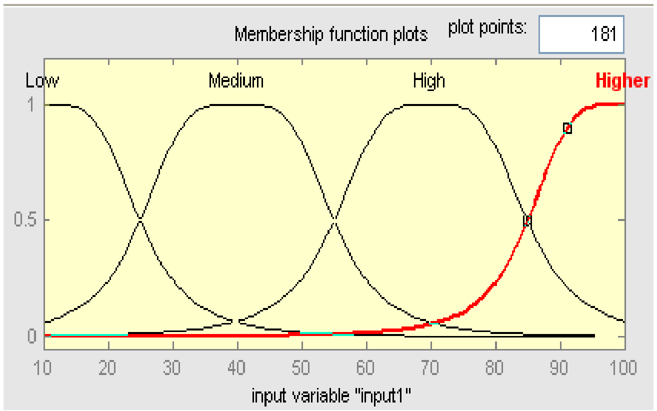
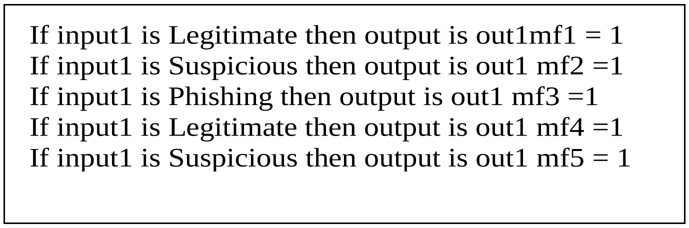


Fig.3. Fuzzy sets values with 4 membership functions after training.



If input1 is Legitimate then output is out1mf1 = 1
If input1 is Suspicious then output is out1 mf2 =1
If input1 is Phishing then output is out1 mf3 =1
If input1 is Legitimate then output is out1 mf4 =1
If input1 is Suspicious then output is out1 mf5 = 1

Fig.4. Rule base containing 5 fuzzy IF-THEN rules

E. Testing Framework

After the training was completed, the checking set was used to check and to test the model. The training process is repeated twice and the testing process is also repeated two-times utilizing training and validation sets only once.

The results are observed. Training outputs are presented in Fig. 3 which is input membership function, type generalized bell shape (Gbell) membership function with the value range of [0, 1] in Y-axis and a value range between [10, 100] on the X-axis. It is defined by linguistic terms including: low indicating legitimate, medium as presents suspicious, while high indicates phishing.

F. Basic Rules

Fuzzy IF-THEN rules are expressed in the form:

If A Then B, where A and B are labels of fuzzy sets [29] characterized by appropriate membership functions. Regarding their concise form, fuzzy if-then rules are usually utilized to obtain the imprecise modes of reasoning that does an important role in the human ability to decide in an environment of uncertainty and imprecision. A description of a simple fact in phishing detection is: If the risk is high or 100% risk, then it is a phishing. If the risk is 0% risk then it is a legitimate. Any number of risks between 0% to 100 is suspicious. An example of rules is shown in Fig. 4. During training, the learning algorithms learn data and use it to create rules. If-Then rules are used because fuzzy rules have been widely utilized successfully in controls and modeling [15].

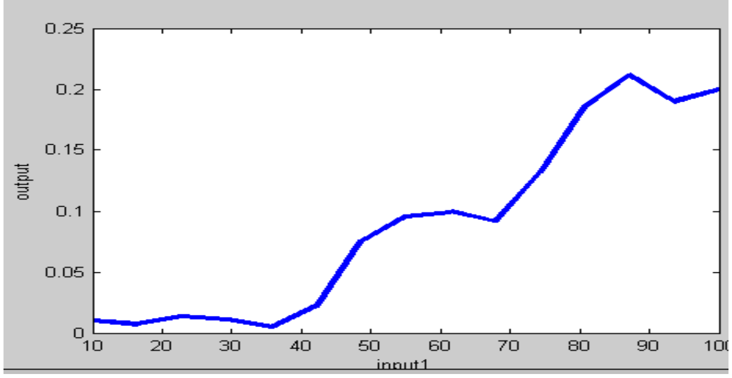


Fig.5. Performance evaluation graph for phishing website detection

TABLE.III. TRAINING AND TESTING RESULTS

Training set	validation set	Training Average error %	Testing Average error %	Testing Error Result %	Average Accuracy Results
150	150	0.012643	0.0126431	1.3%	98.74%

V. TESTING RESULTS AND DISCUSSIONS

After conducting extensive experiments, testing results are obtained in average error rates, which is a measure of the model accuracy performance in real time. The exact measurement is the overall output in which the model is compared. Fig. 6 presents the results as follows: Blue crosses on graphs indicate training results, while red stars indicate test results. An average test error rate obtaining is 0.012631 as shows in fig. 6. Fig. 4 also presents the system performance The training and testing errors are converted in to percentages and presented in Table III. Average testing errors in column 4 are rounded to 2 decimal places and converted in to percentage average error rates which is 1.3% as shown in column 5 and error rates into accuracy percentage in which is 98.74% overall achievement.

VI. ANALYSIS

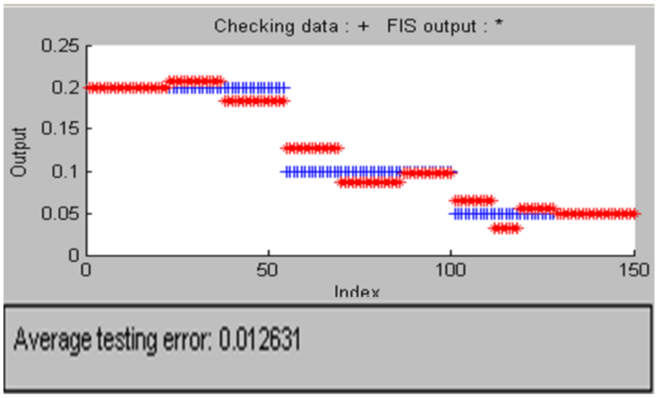
The parameter tuning framework was evaluated using 2-fold cross-validation methods to measure the capability of the model. Parameters were assigned 4 membership functions values, and set to linear the output membership functions, hybrid was assigned as parameter optimization methods. 16 epochs are assigned so that the process stops at 0.01the minimal error tolerance. 0.012631 average errors was obtained, which demonstrated best results compared to other previous works. Our model suffered a modest error rate of 1.6%, which can be explained that the 4 membership function value was not the least visible by the given data. Thus is greater than the given variable example. Otherwise, the lower the average error rates, the better the results. The highest result achieved is nearer to the expected results, given the target performance to be closer to 100% accurate if not 100% accurate. In which case, 98.74% accuracy is nearer enough.

A. *Comparisions*

The techniques and the previous results are compared to determine the best results. The proposed approach utilized 300 data set randomly split in 150 training pair and testing on the remaining set which demonstrated an improvement of 0.34% higher compared to our previous work. Our previous work that is being improved which is Framework 3 and Framework 4 that used 228 and 342 features, assigned values of 15 parameters. 3 and 4 MFs were specified and assigned 12 and 10 Epochs. This experiment achieved 98.4%. Therefore the new approached have significant improvement.

To compare our results with other existing results in the field, our results are not directly comparable with the previous results for the following reasons: Firstly, our work has considered all possible components which are used as inputs in which features are extracted, which include legitimate site rules, user-behaviour profile, PhishTank, user-specific site, pop-up windows and user’s credential. Secondly, from those inputs, 342 comprehensive data are gathered that were used for modeling.

Thirdly, adaptive neuro-fuzzy algorithm has been our proposed work which has not been considered in phishing detection field by other studies in this field. The previous work for example: Aburrous’s studies applied fuzzy logic and datamining techniques with 27 features to detect phishing websites and achieved 83% and 84.4% accuracy [2], [32]. Aburrous’s studies suffered high false positives. They only considered phishTank as their source for 27 features which are a small size. Ma also used a similar approach to Aburrou, but with large lexical features extracted from URL only [5], [28]. They achieved 95-99% accuracy.



Average testing error: 0.012631

Fig.6. Result for framework

These previous studies have not actually used all the possible data in terms of size and diversity, therefore our 98.74% accuracy is much stronger than the existing results. Moreover parameter tuning framework has not been considered in literature in this field [9].

B. Findings

Based on the results of our experiment, we found that applying adaptive neuro-fuzzy algorithm with comprehensive data and proper parameter tuning can detect phishing website with high accuracy. We also found that while data and parameters can have influence on model performance, parameters have direct effect.

C. Limitations

In light of our results from extensive experiment, our results suffered an average errors rate of 1.3%. This can be explained that there was some defective data that caused overfitting and or unrefined parameter tuning also confused parameter that caused the model performance to suffer. The challenge in using ANFIS is that input membership function parameter is limited to either constant of linear.

VII. Conclusions and Future Work

Data has been extracted. Extensive experiments have been conducted. During 10-fold cross-validation data has been randomly split into train and to validate sets. We found that using comprehensive data through ANFIS with proper parameter tuning can detect phishing websites with high accuracy.

A. Contributions

The main contributions made in this paper includes: (1.) identified user's credential profiles as one of core component of input data that has not been utilized in the field. (2.) introduced novel data based on user's credential profile, introduced novel parameter tuning framework based on ANFIS algorithm using comprehensive feature, (3.) applied ANFIS for the first time in (phishing detection website) a new field.

The information about parameter tuning can provide a novel knowledge to researchers about the capabilities of applying ANFIS with comprehensive data and proper parameter settings.

The advantage is that the outcome from this study should provide a great knowledge and understanding to researchers in the field. The method can also be used across other fields in solving similar problems.

B. Feature work

The work do be done next is to extract large data from a wide range of samples and use different cross-validation with large data-sets.

References

[1] Financial Fraud Action UK, Cheque & Credit clearing Company, UKCARDS Association. Deception crimes drive small increase in card fraud and online banking fraud losses. Press Release, pp. 2, 2012 [online] www.financialfraudaction.org.uk. Accessed 24.7.2013.

[2] M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah, "Modelling Intelligent Phishing Detection System for e-Banking using Fuzzy data Mining," International, 2009.

[3] N. Sanglerdsinlapachai, and A. Rungsawang, "Using Domain Top-page Similarity Feature in machine learning-based Web Phishing Detection," In Procedings of IEEE 3rd International Conference on knowledge Discovery and Data Mining, pp. 187-190, 2010.

[4] G. Xiang, B. A. Pendleton, J. Hong, "Modelling content from human-verified blacklist for accurate zero-hour phish detection," probabilistic approach for zero hour phish detection, In Proceedings of the 15th European, 2009.

[5] J. Ma, L. Saul, S. Savag, G. Voelker, "Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs," Proc. of the 15th International Conference on Knowledge Discovery and Data Mining, Paris, France, pp. 1245-1254, 2009..

[6] PhishTank Site Checker (2013), GS! Networks, [online] < https://addons.mozilla.org/en-US/firefox/addon/phishtank-sitechecker/reviews/> Accessed 22.2.2014.

[7] C. Whittaker, B. Ryner, M. Nazif, "Large-Scale Automatic Classification of Phishing Pages," In the 17th Annual Network and Distributed System Security {NDSS'10} Symposium, 2010.

[8] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, C. Zhang, "An empirical analysis of phishing blacklists" in Proceedings of the 6thConference on Email and Anti-Spam, 2009.

[9] J. S. R. Jang, "ANFIS*:* adaptive-network-based fuzzy inference system. IEEE,". Transactions on systems, MAN, and Cybernetics, Vol. 23, No. 3, 1993.

[10] R. Suriya, K. Saravanan, A. Thangavelu, "An integrated approach to detect phishing mail attacks a case study," SIN Proceedings of the 2nd international conference on Security of information and networks, north Cyprus, Turkey, October 2009, 6-10, pp. 193-199, vol. 3 ACM New York.

[11] L. Wenyin, N. Fang, X. Quan, B. Qiu, G. Liu, "Discovering Phishing Target based on Semantic Link Network," *Future Generation Computer Systems*, Elsevier, Volume 26, Issue 3, March 2010, pp. 381-388.

[12] G. Xiang, B. A. Pendleton, J. I. Hong, C. P. Rose, "A hierarchical adaptive," Symposium on Research in Computer Security (ESORICS'10). 268–285, 2010.

[13] X. Dong, J. A. Clerk, J .L. Jacob, "Defending the weakest link: Phishing Website Detection by analysing User Behaviours," IEEE Telecommun System, 45: pp. 215 – 226, 2010.

[14] B. Wardman, T. Stallings, G. Warner, A. Skjellum, "High-Performance Content-Based Phishing Attack Detection," eCrime Researchers Summit (eCrime), pp. 1-9, Conference: 7-9 Nov. 201 1, San Diego, CA.

[15] A. P. Barraclough, M. A. Hossain, M.A. Tahir, G. Sexton, N. Aslam "Intelligent phishing detection and protection scheme for online transactions," Expert Systems with Application 40, pp. 4697-4706, 2013.

[16] A. Le, A. Markopoulou, M. Faloutsos, "Phishdef: Url names say it all," INFOCOM, Proceedings IEEE, pp. 191-195, 2010.

[17] H. Huh, H. Kim, "Phishing Detection with popular search engine: Simple and effective", In Proceeding FPS'11 Proceedings of the 4th Canada-France MITACS conference on Foundations and Practice of Security, pp 194-207, 2012.

[18] D. Canali, M, Cova, G. Vigna, C. Krugel "Prophiler: A fast filter for the large-scale detection of malicious web pages," In Proceedings of the International World Wide Web Conference,. 2011.

[19] G. Xiang, J. Hong, C. P. Rose, L. Cranor," Cantina+: A feature-rich machine learning framework for detecting phishing web sites," ACM Transactions on Information and System Security (TISSEC), 14(2), pp. 2- 21, 2011.

[20] W. Ead, W. Abdelwahed, H. Abdul-Kader, "Adaptive Fuzzy Classification- Rule Algorithm in Detection Malicious Web Sites from Suspicious URLS," International Arab Journal of e-Technology 3 (1), pp. 1-9, 2013.

[21] G. Xiang , "Toward a phish free world: A feature-type-aware cascaded learning framework for phish detection", Thesis, Language technologies institute, School of computer science., 2013.

[22] Pinsent manson law expert, (2011). [online] http://www.pinsentmasons.com/en/expertise/sectors/core-industries--markets/universitiesandhighereducation/ Accessed 28.12.11.

[23] PhishTank, "Join the fight against phishing," 2011. [online] < http://www.phishtank.com/ > Accessed 5.6.2012 and 10.7.2013.

[24] Barclays Bank "online banking," 2012. [online] < http://www.barclays.co.uk/ > Accessed 8.12.2012.

[25] Financial Service Authority (FSA), (2013), UK [online] http://hb.betterregulation.com/external/List%20of%20banks%20-%2028%20February%202013.pdf and <www.fsa.gov.uk> Accessed 8.12.2012.

[26] G. B., Huange, Q. Y., Zhu, K. Z., Mao, C. K., Siew, P. Saratchandran, & N.Sundararajan, , (2006). Can threshold networks be trained directly*? IEEE Trans. Circuits syst. II*, vol. 53, no 3, 187-191.

[27] R. Kohavi, (1995). A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *The International Joint Conference on Artificial Intelligence*, Computer Science Department Stanford University (IJCAI).

[28] J. T. Ma, "Learning to detect malicious URLs," Thesis, University of California, 2010.

[29] P. B. Sivarao, N.S.M. El-Tayeb, "A New Approach of Adaptive Network-Based Fuzzy Inference System Modelling in Laser Processing- A Graphical User Interface (GUI) Based," Journal of Computer Science. 5 (10), pp. 704-710, 2009.

[30] Complying with anti-phishing regulation (2012) http://help.wildapricot.com/display/DOC/Complying+with+anti-spam+regulations

[31] PhishTank, "Join the fight against phishing," 2012. [online] < http://www.phishtank.com/ > Accessed 5.7.2013 and 10.7.2013.

[32] M. Aburrous, M. A. Hossain, K. Dahal and F. Thabtah "Intelligent phishing detection system for e-banking using fuzzy data mining,"Expert Systems with Applications 37, pp. 7913-7921, 2010.

APPENDIX

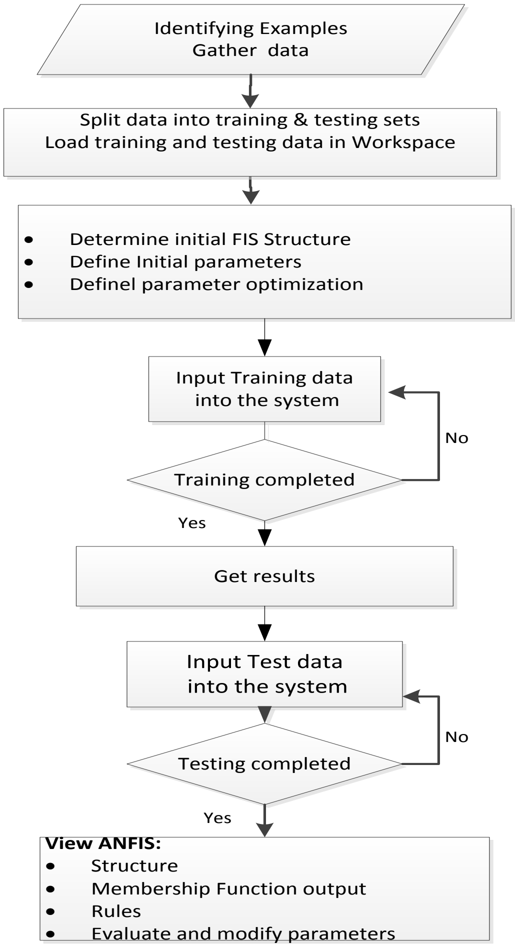

Identifying Examples
Gather data
Split data into training & testing sets
Load training and testing data in Workspace
Determine initial FIS Structure
Define Initial parameters
Definel parameter optimization
Input Training data into the system
No
Training completed
Yes
Get results
Input Test data into the system
No
Testing completed
Yes
View ANFIS:
Structure
Membership Function output
Rules
Evaluate and modify parameters

Fig.7. Conceptual method

FlexRFID: A Security and Service Control Policy-Based Middleware for Context-Aware Pervasive Computing

Healthcare Scenarios

*Mehdia Ajana El Khaddar[1], Mhammed Chraibi[2], Hamid Harroud[3], Mohammed Boulmalf[4], Mohammed Elkoutbi[1], Abdelilah Maach[2]

1: SIME Lab, ENSIAS, Rabat, Morocco
2: Ecole Mohammedia des Ingénieurs, Rabat, Morocco
3: Alakhawayn University in Ifrane (AUI), Ifrane, Morocco
4: International University of Rabat (UIR), Rabat, Morocco

***Abstract*— Ubiquitous computing targets the provision of seamless services and applications by providing an environment that involves a variety of devices having different capabilities. The design of applications in these environments needs to consider the heterogeneous devices, applications preferences, and rapidly changing contexts. RFID and WSN technologies are widely used in today's ubiquitous computing. In Wireless Sensor Networks, sensor nodes sense the physical environment and send the sensed data to the sink by multi-hops. WSN are used in many applications such as military and environment monitoring. In Radio Frequency Identification, a unique ID is assigned to a RFID tag which is associated with a real world object. RFID applications cover many areas such as Supply Chain Management (SCM), healthcare, library management, automatic toll collection, etc. The integration of both technologies will bring many advantages in the future of ubiquitous computing, through the provision of real-world tracking and context information about the objects. This will increase considerably the automation of an information system. In order to process the large volume of data captured by sensors and RFID readers in real time, a middleware solution is needed. This middleware should be designed in a way to allow the aggregation, filtering and grouping of the data captured by the hardware devices before sending them to the backend applications. In this paper we demonstrate how our middleware solution called FlexRFID handles large amount of RFID and sensor scan data, and executes applications' business rules in real time through its policy-based Business Rules layer. The FlexRFID middleware provides easy addition and removal of hardware devices that capture data, as well as uses the business rules of the applications to control all its services. We demonstrate how the middleware controls some defined healthcare scenarios, and deals with the access control security concern to sensitive healthcare data through the use of policies. We propose hereafter the design of FlexRFID middleware along with its evaluation results.**

Keywords*— RFID; Middleware; WSN; Ubiquitous; Pervasive Computing; FlexRFID; Policy-Based; Security; Healthcare; access control

I. Introduction

Pervasive computing aims at providing intuitive and seamless support for the users through leveraging the distinct functionalities of a number of devices, and developing various backend applications that use data gathered from these devices. Through wireless communication, the automation devices can share data, and combine them for a more accurate inference of their surroundings. This inference enables applications to reason about the past, the present, and the future, and allows them to behave according to the expectations of the user. This is making pervasive applications very attractive to users on one hand and close to nightmare for developers on the other hand. This is due to the fact that pervasive applications need to deal with device heterogeneity, unreliable wireless communication, duplicate and continuous raw data readings, uncertainty in sensor readings, and changing user requirements and application domains. Therefore, the development of this kind of applications is considered error prone, non-trivial and time consuming, and needs definitely a rescue which is a middleware for pervasive computing [25].

Healthcare services are becoming increasingly pervasive where monitoring technologies are fast becoming integral to the care process and important to realize a proficient healthcare service. WSN and RFID can be considered two adjacent technologies that help tracking healthcare items and patients, and providing context information about them. Sensors measure physiological state (inpatient monitoring), and also allow remote care (outpatient monitoring, i.e. at the patient's home rather than in hospital) [1]. RFID not only offers tracking capability to locate patients in real time while they are moving in a hospital, but also monitors access control to the different medical departments, and provides efficient and accurate access to medical data for doctors and other health professionals [2]. Such technologies assist in the early identification of health issues, and provide alerts in case of emergencies.

The healthcare environment is becoming data driven, in the sense that care providers require information in order to deliver care services. However, health information is sensitive and must be protected [3]. Thus, it is necessary to consider the context in which it is shared. The development of a middleware, which hides the complexity of the underlying network and eases application development, is central to provide a ubiquitous secure healthcare.

The solution proposed in this paper is a middleware which supports simultaneous communication of multiple applications with the RFID and WSN hardware, and deals with the above challenges through the use of policies. The middleware provides all data processing capabilities like filtering, grouping and duplicate removal. The paper is structured as follows, Section II introduces related work. Section III introduces the middleware architecture and focuses more on the policy-based Business Rules Layer, presents the policies types and structure, and shows how all the services provided by FlexRFID can be managed by the use of application-defined policies. Section IV defines and models policies for some healthcare scenarios, followed by conclusions and future work in section V.

II. Related Work

Most of the existing RFID middleware solutions are commercial. These include “BizTalk RFID” middleware from Microsoft, and “Java RFID System” from Sun, to name a few. Other middleware solutions were developed from research e.g. “WinRFID” by UCLA and “Accada” by ETH Zurich. Sun Java RFID System is a Java based commercial middleware that has a dynamic service provisioning architecture that enables scaling from small to large deployments with high data volume [4]. The Biztalk RFID middleware solution from Microsoft provides support for both standard and non-standard devices through the plug-and-play architecture [5]. It has an event processing engine that manages the RFID events by creating business rules, through which it provides real time visibility of the RFID data [5]. WinRFID [6] developed at the University of California Los Angeles (UCLA), uses web services and enables rapid RFID applications development. It has certain unique features like hiding of communication details from the end users, network management on a large scale, intelligent data processing and routing, support for hardware and software interoperability, provision for system integration and system extendibility, etc. WinRFID exploits the .Net framework‘s runtime plug-in feature to support the addition of new readers, protocols, and data transformation rules with minimum disruption of the existing infrastructure [6]. The Accada middleware [7] developed by ETH Zurich uses EPCglobal (Electronic Product Code) based specifications for the reader protocols, the application level event specifications and the EPCIS (EPC Information Services) capture and query interface to handle RFID data flow across enterprises. It has three main modules: the reader, the middleware, and the EPC information services module. The Accada reader implementation uses standard edition of SUN Java Virtual Machine [7].

For WSN, there exists many middleware approaches. Among these approaches we find the virtual machine, database, application driven, message-oriented, and modular programming middleware [8]. For each of the WSN middleware approaches, some WSN middleware solutions have already been proposed. Hereafter we name some WSN middleware solutions: Impala [8, 9], Mate [10, 8], Middleware Linking Applications and Networks (MiLAN) [8], Sensor Information and Networking Architecture (SINA) [8], Mires [8], etc.

A publish/subscribe middleware providing event based data control mechanisms for healthcare has been proposed in [11]. The main objective of this event based healthcare middleware is to give caregivers fine-grained control over the circumstances for health data transmission. It provides two categories of interaction control rules; subscription rules and event transformation rules. These rules allow the administrative domain to set the circumstances in which it is appropriate to transmit particular information.

A middleware architecture using the MVC pattern is proposed in [12]. The proposed middleware architecture for pervasive computing supports the architectural quality attributes of adaptability, availability, security, and modifiability. All these requirements are ensured using the MVC design pattern as discussed in [12].

Fusion from ORACLE is a Service Oriented Architecture (SOA) middleware for healthcare integration, connecting administrative and clinical processes. Through the use of Fusion Middleware, Oracle helps organizations to reliably exchange information while adhering to important industry standards and initiatives. This enables organizations to lower operating costs and accelerate time-to-market by delivering a consistent user interface, security architecture, management console, and monitoring environment. The Fusion middleware is designed to correlate clinical data, link applications, and comply with the myriad challenges of this highly regulated, data-intensive industry. Smoothing data interchange helps streamline every phase of the healthcare lifecycle from initiation, eligibility, and enrollment to service delivery, program analysis, and reporting [13].

There exist many other pervasive computing middleware solutions e.g. *SeSCO*, *OneWorld*, and *AoC* to name a few [22]. Though the existing middleware solutions are useful, they themselves have varied features and contribute partially, to context, data, or service management related application developments. Most of them are oriented toward mobile applications and do not provide abstraction for many types of applications. There is no single middleware solution that can address a majority of pervasive computing application development issues, due to the diverse underlying challenges. Also, there is a huge scope for research in the area of RFID and WSN middleware and applications, and many solutions were developed to cope with the technology –related challenges.

As compared to the related work described herewith, FlexRFID middleware has many distinguishing aspects. It provides the applications with a device neutral interface to communicate simultaneously with many different hardware devices, creating an intelligent network of RFIDs, sensors, and any other type of automation devices. The policy-based Business Rules Layer allows FlexRFID middleware to enforce all data processing capabilities by applying the backend applications defined rules. This enforces security by restricting access to data only to the applications or users that satisfy certain conditions stated in the policies. Also the modular layer of the middleware allows seamless integration of different types of enterprise applications, which makes it a general middleware not related only to one application domain. We present hereafter the middleware architecture, and show how business rules are modeled and applied by the middleware using some healthcare scenarios.

III. PROPOSED POLICY BASED MIDDLEWARE AND ITS SUBSYSTEMS

A. FlexRFID Middleware Overview

The FlexRFID middleware as described in [14] is a multi-layered middleware consisting of Device Abstraction Layer (DAL) which abstracts the interaction with the physical network devices, Business Event and Data Processing Layer (BEDPL) which provides data services such as dissemination, aggregation, transformation, and duplicate removal, Business Rules Layer (BRL) which is a policy-based management engine that defines the rules that control resources and services of the FlexRFID middleware, and Application Abstraction Layer (AAL) which provides a high level of software abstraction that allows communication among the enterprise applications and the FlexRFID middleware. FlexRFID was integrated in many domains: library management [14], inventory control with Opentaps software [15, 16], and healthcare [17].

The BRL is a policy-based management engine that defines the rules that grant or deny access to resources and services of the FlexRFID middleware, and enforces different types of policies for filtering, aggregation, duplicate removal, privacy, and different other services. This is achieved by determining the policies to apply when an application requests the use of a service in the BEDPL. Hereafter, we give more details about the BRL, policies architecture, and examples of policies representation.

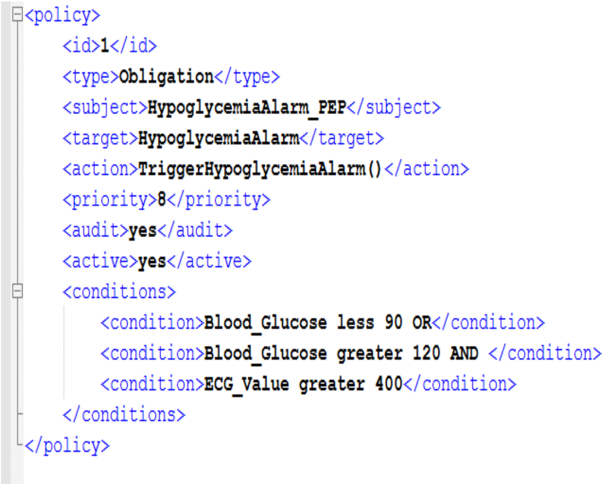
B. Policy-Based Business Rules Management Layer

1) Policy Types and Structure

Software policies have been widely used to provide security for WSN as in SecSNMP [18]. Policies are operating rules used to maintain order, security, consistency, or other ways of successfully achieving a task. In our middleware architecture we are using the types and structure of policies defined for the system in [19]. There are basically two types of software policies. *Authorization policies* are rules that are usually enforced in access control systems. In the case of the FlexRFID middleware, authorization policies would be rules defined by the application to enforce or deny access to certain data if a certain set of conditions is fulfilled. *Obligation policies* refer to actions to be enforced when a set of predefined conditions is fulfilled or a change in the context happens. For FlexRFID middleware, an obligation policy would be to trigger the duplicate removal service for an application only 5 minutes of data reads. The policy specification language, based on which we have decided to model our policies is Ponder, and we have chosen XML to represent policies due to its ease of editing and use.

As presented in [19], the policy has nine main attributes. The *policy ID* is a unique identifier of the policy and helps in the search operation. The *type* of policies refers to whether we deal with an authorization or obligation policy. The *subject* of the policy is the entity that enforces the action of the policy, and the *target* is the entity on which the policy's action is enforced. Usually the action of the policy is a call for a method that belongs to the target. The *priority* attribute of the policy is used to solve the problem of conflicting policies. The *audit tag* of the policy allows the system to keep track of the triggered policies and their context, and the *active tag* specifies if a policy is active or not. One of the most important attributes of the policy is the *set of conditions* which refers to conditions under which the policy is triggered. These are expressed using first order logic and comparison operators. Context information that is included in the policies is part of the condition set. For example this context information could refer to time, location, type of application, and role of users.

Fig. 1 below shows an example policy for blood glucose management, with the attributes mentioned above. In this example, policy with ID 1 is an obligation policy that triggers an alarm for a specific group of physicians taking care of a diabetic patient facing a hypoglycemia state. When the blood glucose (Blood_Glucose) and heart rate (ECG_Value) values are communicated by the sensors, they are checked to see whether they satisfy the set of conditions mentioned in the policy.



```
<policy>
    <id>1</id>
    <type>Obligation</type>
    <subject>HypoglycemiaAlarm_PEP</subject>
    <target>HypoglycemiaAlarm</target>
    <action>TriggerHypoglycemiaAlarm()</action>
    <priority>8</priority>
    <audit>yes</audit>
    <active>yes</active>
    <conditions>
        <condition>Blood_Glucose less 90 OR</condition>
        <condition>Blood_Glucose greater 120 AND </condition>
        <condition>ECG_Value greater 400</condition>
    </conditions>
</policy>
```

Fig. 1.   Sample policy structure for hypoglycemia management

The policy in Fig. 1 states that If the condition set is met, the Trigger_HypoGlycemiaAlarm() method of the doctors' application is called to inform them about this emergency case.

2) Policy-Based BRL Architecture

In general a policy management system is composed of three main entities; the PDP (Policy Decision Point), the PEP (Policy Enforcement Point), and the PIB (Policy Information Base). The PDP is responsible for taking the decision whether to allow or deny an action based on the request details and the policies available in the PIB. The PIB refers to the database containing all the system policies. Once the action is processed and selected by the PDP, it sends a message to the PEP which is responsible for enforcing the action on the target. In our system we kept the same components and added some others that ease management of policies and adapt the policy management system to the middleware.

Fig. 2 shows the architecture of the FlexRFID middleware adapted to put more emphasis on the components of the policy-based BRL, and explain interaction between them.

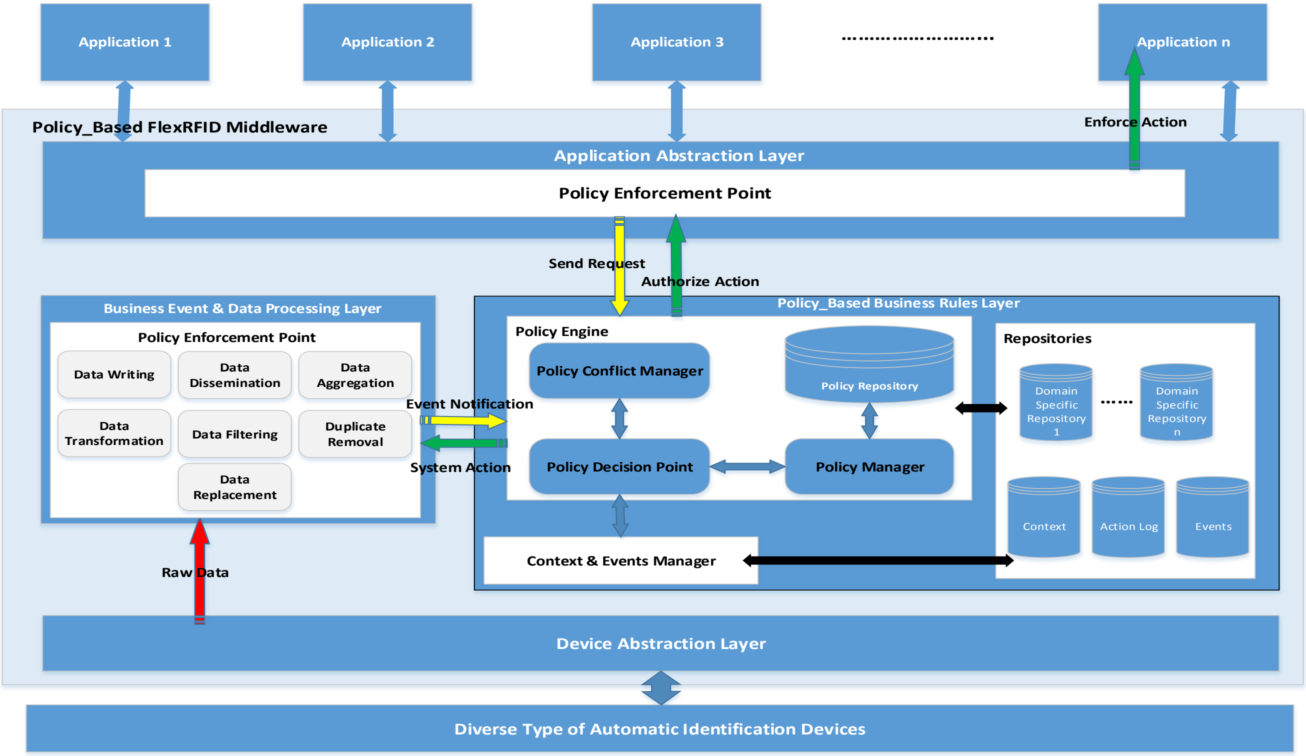


Fig. 2. Policy-based Business Rules Layer (BRL) architecture of the FlexRFID middleware

The BRL contains three main components: the Policy Engine, the Context and Events Manager, and the Repositories. The *policy engine* is responsible for policy management of the middleware and has four main components namely the PDP, the policy manager, the policy conflict manager, and the policy repository. The *policy manager* accesses the policy repository to read the policies and organize them for use by the two other components; the PDP and the policy conflict manager. The policy manager is also responsible for updating the policy repository by adding new policies when needed by the applications or when new applications subscribe to the middleware. The *policy conflict manager* sorts the list of policies given by the PDP in increasing order of priority, and triggers the policy with the highest priority for a specific event. The PDP is responsible for evaluating the policies and deciding whether a policy action is to be triggered or not. The PDP is either triggered by an incoming request or event that is external to the system for example new sensor data, or by an internal event that is a notification from the context manager of a change in the environment's context. The *policy repository* contains all the obligation and authorization policies of the system.

There are four main repositories in the BRL. The *context repository* contains all context information that is of use to our system such as time, and location and is used by the context manager. The *actions log repository* contains a log for every policy that has been triggered. The log helps providing with accountability such as the requester identification, information about the type of the policy; whether it is an obligation or an authorization policy, the subject and the target. The *events repository* contains all the events encountered by the middleware for example the read of a new RFID tag, or new sensor data detection, a change in the object location, etc. The *domain specific repositories* contain data related to a certain application domain. In the case of healthcare the domain specific repository would be the *Electronic Health Record* (EHR).

The PEP stands both at the AAL and BEDPL of the middleware because it is the one responsible for performing actions that are specified in the system and applications policies. At registration phase with the middleware, the application loads its set of policies to the middleware so that both PEPs can enforce actions specified in the policies when the condition set is met.

The sequence diagram in Fig. 3 shows how the different components of the policy based BRL interact with each other and with the remaining middleware components when new data is detected. Once the automatic identification device detects the data, they are sent to the service PEP at the level of the BEDPL. The PEP in its turn sends the event to the PDP.

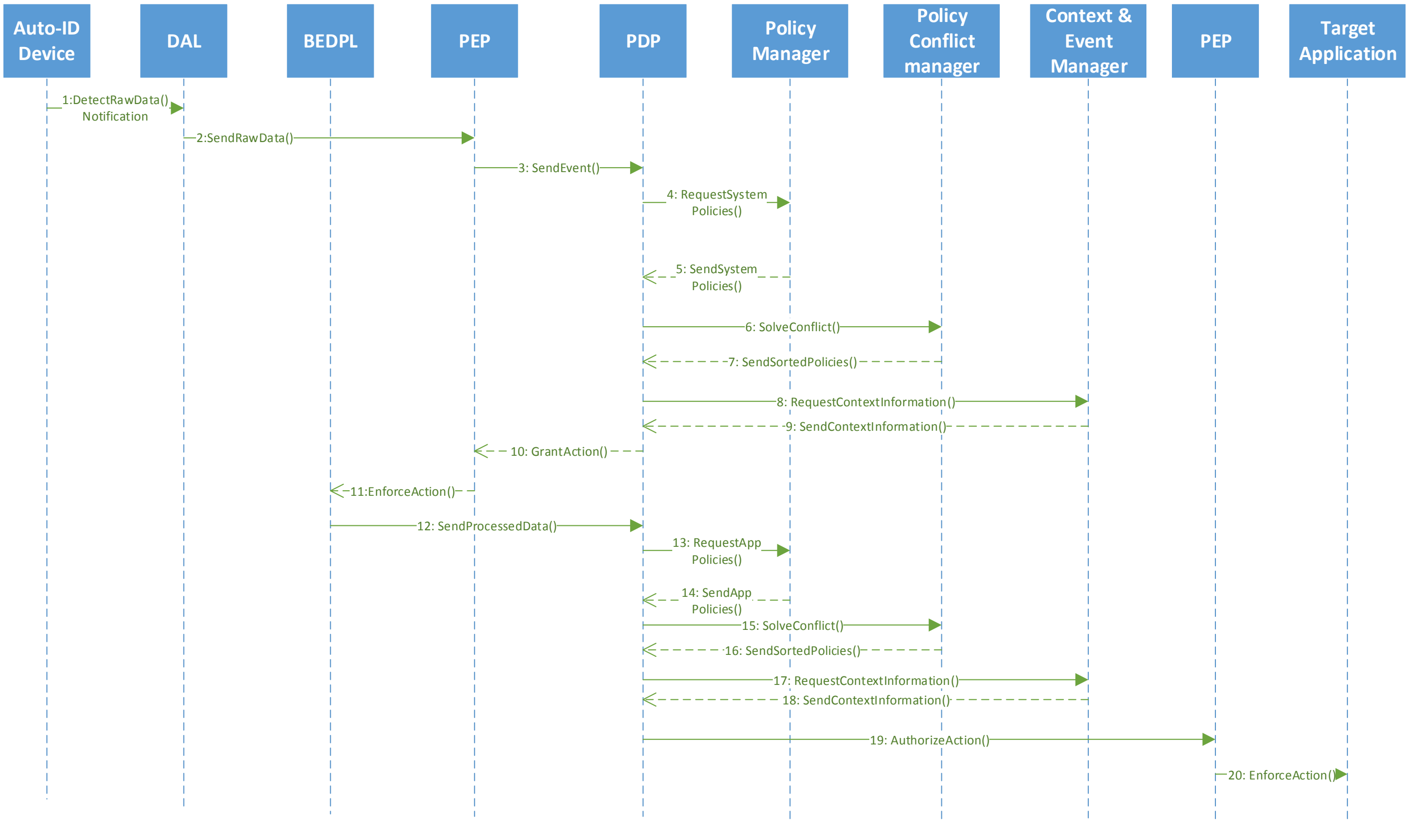


Fig. 3. Policy-based data processing flow in FlexRFID middleware

The PDP requests system policies in charge of controlling the BEDPL services from the policy manager, and requests solving conflict from the policy conflict manager in case of conflicting policies. PDP also requests the context information from the context manager if available. After resolving conflict and acquiring context information the PDP grants the action to the service PEP which in turn enforces the target service at the level of BEDPL. The BEDPL sends these processed data to the PDP for application of further policies related to the backend application business rules. The PDP checks again the application policies, solves conflict, and requests context information if needed. Once the policies are applied it authorizes the action for the PEP at the application level, and the PEP enforces the action on the target application. As can be seen from the sequence diagram in Fig. 3 policies are divided logically into two types: system policies related to the middleware services, and application related policies which refer to any additional rules that need to be applied to data after the first processing by the BEDPL.

IV. HEALTHCARE SCENARIOS

Healthcare is a highly collaborative environment which requires information sharing to provide care. Different services take part of the care process; a patient gets admitted to a hospital, doctors prescribe treatment for their assigned patients, prescriptions are sent to pharmacy, and specific information is sent to accountants for the billing purpose.

With the provision of sensor and automatic identification technologies, we are more talking about *homecare* and *remote care* outside of traditional care institutions. This reduces the need for human intervention, alerts of particular incidents, and provides detailed representations of patient state in real-time. The patient becomes empowered and feels more independent while receiving more information to assist him/her in self-care [20].

Remote care environments are dynamic and each instance of them is created on demand to cater for specific aspects of patient care. The instance is customized to a particular situation in terms of management policies and entities involved. Remote care requires entities that deliver services depending on their role in the care process, and requiring notification of events as they occur. Events include either actions performed such as a patient taking a drug, data notification such as a sensor monitoring a vital sign, and state changes such as a detection of an emergency case.

In a healthcare application, the FlexRFID middleware should support the real-time dissemination of events to the interested entities, while providing support for heterogeneous devices capturing data and means to control information processing and disclosure. This involves loading the policies into the middleware by the different healthcare entities or applications in order to define the situations for data release.

Example of policies in healthcare may include the following:

Subscription for particular events: this policy is used in case a user may request particular information for delivery as it occurs. For example a policy might allow a doctor to receive treatment events only for patients that he treats.

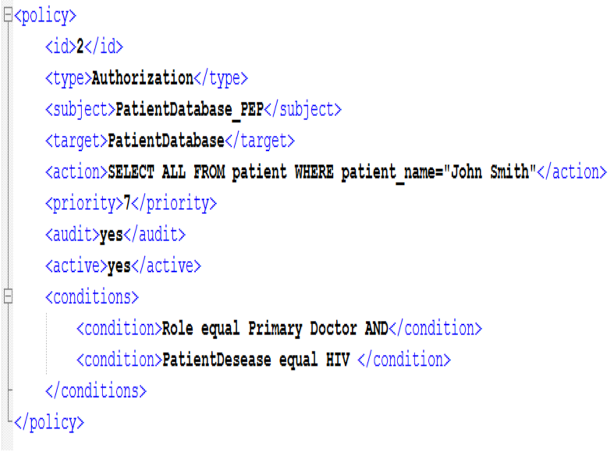
Data access control and event restriction: this policy defines the conditions under which certain data or events are not delivered for a subscribed application. For example a policy might prevent a nurse from receiving a patient's HIV treatment data, while allowing this for his caring doctor. Another policy might allow a physician to modify any medical record for which he or she is designated as primary physician. Also a policy can state that access to a medical record is allowed for five times only and each access expires after one minute.

Data and event transformation: this policy involves modifying the event type or attributes to better satisfy the requirements of the subscribed entity or its current context. It is a context-aware policy that accounts for emergency cases. For example if a doctor has no relationship with a patient, their subscription can be deactivated. Another example is to transform glucose reading to an alert if it is too high for two successive days.

Hereafter we describe the application of our policy-based data control middleware to two healthcare scenarios.

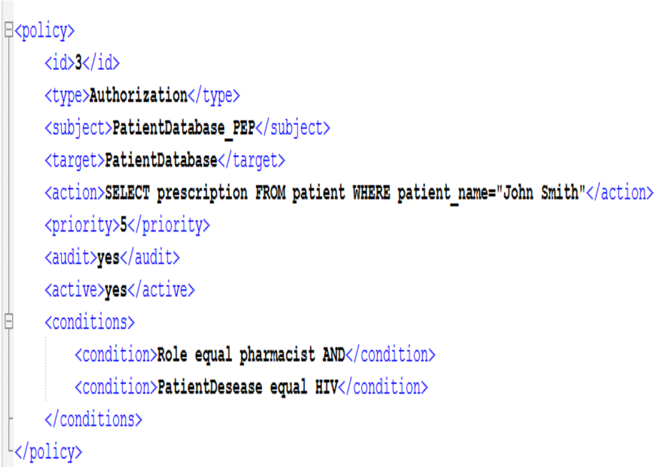
A. Drug prescription control scenario

A nurse may prescribe some controlled drugs in certain circumstances to a patient. This prescription must be validated by the primary treating doctor, who must access to all details about the patient care and history. The prescription must flow to the pharmacy without any details of the patient and the reason for prescription. An auditor must monitor the supply of the controlled drug and must not receive patient specifics. This emphasizes role based access control to patient data. Fig. 4 and Fig. 5 below show policies used for the above scenario to control drug prescription.



```
<policy>
    <id>2</id>
    <type>Authorization</type>
    <subject>PatientDatabase_PEP</subject>
    <target>PatientDatabase</target>
    <action>SELECT ALL FROM patient WHERE patient_name="John Smith"</action>
    <priority>7</priority>
    <audit>yes</audit>
    <active>yes</active>
    <conditions>
        <condition>Role equal Primary Doctor AND</condition>
        <condition>PatientDesease equal HIV </condition>
    </conditions>
</policy>
```

Fig. 4. Data access control policy to patient's data by the primary doctor



```
<policy>
    <id>3</id>
    <type>Authorization</type>
    <subject>PatientDatabase_PEP</subject>
    <target>PatientDatabase</target>
    <action>SELECT prescription FROM patient WHERE patient_name="John Smith"</action>
    <priority>5</priority>
    <audit>yes</audit>
    <active>yes</active>
    <conditions>
        <condition>Role equal pharmacist AND</condition>
        <condition>PatientDesease equal HIV</condition>
    </conditions>
</policy>
```

Fig. 5. Data access control policy to patient's data by the pharmacist

B. Location tracking and emergency management scenario

Location sensors or RFID tags are commonly used in remote care to detect the location of patient. Tracking location is important in detecting emergencies especially for elderly care [21] (are patients in bed? Did they fall on the floor?), and in quickly dispatching the ambulance to the target location. Patients generally care about privacy and want their exact location to be obscured. A policy might specify that location information should not be transmitted unless it is an emergency situation, and for some defined entities in the care process for example for doctors and close relatives only.

Referring to the hypoglycemia policy modeled in Fig. 1, diabetes self-management can be easily deployed using policies. Sensors can be used to take the diabetic patient measurements like blood glucose, blood pressure, amount of meals, amount of exercise, and location. The measurements taken by the sensors can be sent to the middleware, aggregated, checked for the specified thresholds set by the doctors in the diabetes management application policies, and sent to the specialized doctor in real-time so that he/she can send advice to the patient. If the blood glucose and pressure are noticed to be too high over a certain period of time (a day for example), the doctor may advice the patient to go for a workout, or to take an additional insulin injection to lower the blood glucose. In case of hypoglycemia the doctor should advice the patient to stay in bed and eat more in order to increase his/her blood glucose. In this case if the blood glucose readings are too low and the RFID tag attached to the patient sends location change information, the middleware should trigger an alarm to the doctor so that he sends a dedicated medical team to take care of the situation, because the patient may fall down while moving in a hypoglycemic state.

The examples above highlight key features of policy based management in the middleware. Information is released depending on the context and situation only to the interested and eligible entities. An event can be interesting to many entities but its visibility is controlled by the policies defined beforehand. This ensures security and privacy concerns.

V. FLEXRFID MIDDLEWARE EVALUATION

A. Device Evaluation

Device evaluation corresponds to the portability metric of ISO/IEC 9126 standard [23] in terms of multiple devices support. This metric evaluates the heterogeneity, and scalability of the middleware as the number of devices increase. We mean by scalability here the durability of stable status of the middleware when certain conditions are met, in this case when the number of devices increases.

FlexRFID middleware provides the RFID applications an interface to RFID hardware and other sensors and automatic identification devices, called "*Device Abstraction Layer (DAL)*". Generally the hardware devices are accessed by a set of APIs provided by the device manufacturer, which are specific to each device. In our implementation of the DAL we used wrappers for the manufacturer provided reader APIs, in order to make the reader accessible through FlexRFID. These wrappers call the reader specific API to implement the desired functionality.

The FlexRFID middleware was tested with *Intermec Fixed IF61 reader* which was available in our lab. This reader‘s DLL was loaded to the FlexRFID. After set up of each connection to the IF61 reader, a handle on it is used to support all further communication with this reader. Further devices' DLLs must be identified and added to the FlexRFID, in order for the middleware to support communication with them.

B. Application Evaluation

Application evaluation corresponds also to the portability metric of ISO/IEC 9126 standard in terms of heterogeneous system and application support. It evaluates the level of abstraction of the middleware in terms of providing standard APIs to communicate with multiple backend applications, and also scalability of the middleware as the number of connected applications, the number of policies loaded by the applications, and the number of requests from the applications increases.

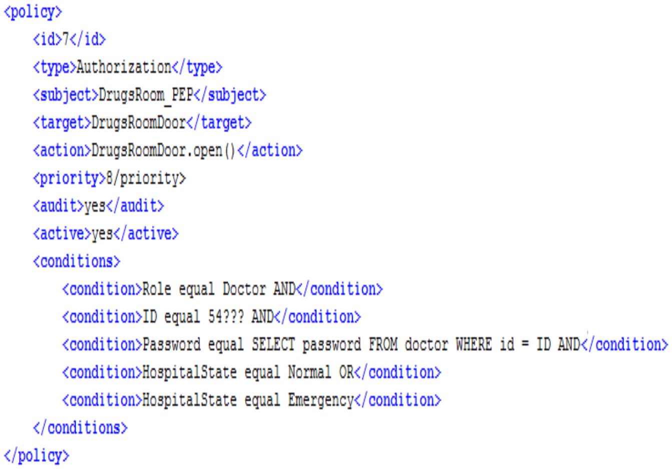
FlexRFID provides through the *Application Abstraction Layer (AAL)* a generic class that should be implemented by all applications that want to connect to the middleware. This class provides functions to access the general operations done by all RFID applications such as reading / writing data, duplicate removal, first level filtering, etc. Domain specific data treatment such as data transformation to some complex business events is either expressed through policies so that the middleware applies the application rules on data before dissemination, or done at the level of the application itself.

FlexRFID was tested with a smart library application prototype [24], integration with OpenTaps for inventory control [16], and we have identified scenarios for healthcare domain integration that we need to simulate in a healthcare application prototype [17].

C. Security and Privacy Evaluation

The security and privacy evaluation is meant to assess the security and privacy of the pervasive middleware, and how it protects the applications' sensitive data when needed through the use of policies. This is achieved by generating a scenario in which access to data is restricted to specific parties, and testing how the middleware deals with this access control policy.

The policy example that we have used in our scenario is shown in Fig. 6 below. This policy maps a business rule from a hospital that says that only doctors who have ID starting with "54" and who have correctly authenticated (ID + Password authentication) can access the "DrugsRoom". The policy restricts access to that specific room to a certain category of hospital employees. The policy enforces the restriction by allowing use of context information such as the role of the employee (Role Based Access Control), and the state of the hospital (Normal state or emergency state). In our test we have generated tags which lead to the creation of requests. The requests contained different specificities which did not match the conditions of the policy responsible for the management of the door leading to the "DrugsRoom". Therefore, none of the requests led to the opening of the door. It is to specify that we have conducted our tests on a local host where no external threats exist.



```
<policy>
    <id>7</id>
    <type>Authorization</type>
    <subject>DrugsRoom_PEP</subject>
    <target>DrugsRoomDoor</target>
    <action>DrugsRoomDoor.open()</action>
    <priority>8/priority>
    <audit>yes</audit>
    <active>yes</active>
    <conditions>
        <condition>Role equal Doctor AND</condition>
        <condition>ID equal 54??? AND</condition>
        <condition>Password equal SELECT password FROM doctor WHERE id = ID AND</condition>
        <condition>HospitalState equal Normal OR</condition>
        <condition>HospitalState equal Emergency</condition>
    </conditions>
</policy>
```

Fig. 6. Healthcare scenario for access control policy.

D. Context Evaluation

The context evaluation assigns metrics to the contexts available in the environment in which the FlexRFID middleware is tested. The context can be time, location, user activities, objects movements, etc. The context evaluation can take place by generating different scenarios by the user and checking the application for context-awareness, for example whether it adapts to changes in context seamlessly. Location tracking and emergency management in healthcare can be a great scenario for showing the middleware's adaptability to context.

Location sensors or RFID tags are commonly used in remote care to detect the location of patient and measure vital signs like blood pressure, temperature, blood glucose, etc. Using policies, in an emergency case, Information is released depending on the context and situation, and only to the interested and eligible entities. An event can be interesting to many entities but its visibility is controlled by the policies defined beforehand. This ensures as well the security and privacy concerns.

E. Performance Evaluation

Response time is among the most important performance parameters. It is the amount of time beginning upon sending some request to the middleware that performs required operations over the massive data gathered from the devices, to receive response from the middleware and disseminate the processed data to the interested applications. Response time measures the delay of query results.

We are interested in two different performance testing cases. First, we have the case where the system is dedicated to one client application. This means that all the policies in the system belong to the same application. Our measure is the time necessary for a request to receive a response depending on the number of policies specified by the client. In the second test, we have decided to have a fixed total number of policies. However, the policies would belong to different client applications. Since we are using one request to test for performance, we wanted to see how the total number of policies affects the system performance. We are in the process of testing these two test cases and getting the performance results.

VI. Conclusion and Future Work

Policy based FlexRFID middleware was developed to cater to the need of many applications that need to take advantage of the ubiquitous computing technologies in order to automate some data processing. We have outlined the need for using this middleware solution for remote healthcare and the flexibility it offers for handling different scenarios thanks to its policy based Business Rules Layer.

We have implemented a working prototype of the middleware including the policy engine and context management system. The next step is to choose one of the defined scenarios, generate the corresponding events, load the defined policies in the middleware, and see how it handles the data related services, access control to data, and ensures patients' privacy. A further testing of the middleware will include performance and response time testing for the test cases defined above. For example we can see how the middleware behaves when the number of policies loaded by a specific application increases, or when the services need to gather data from many devices for processing, etc.

The current version of the FlexRFID middleware offers only basic support for security through the use of access control policies because the project intentionally focused on the described concepts and policies definition. More work on the security features is needed in the future for e.g. security at the level of tags or sensor nodes, application authentication, reader/tag authentication implementation, etc. Ongoing work will also consider the autonomy evaluation which deals with the following parameters: self-configuration, resource management, failure tolerance, high availability, and decision making. The FlexRFID will also be integrated in the cloud, provide a foundation for enabling applications to flexibly use services provided in the cloud, and automatically adapt the usage of cloud-based services depending on application policies and context considerations. As the technology matures in the future, FlexRFID may integrate other types of devices, and handle new services and applications.

References

[1] M. Amimian, and H. R. Naji, "A hospital health care monitoring system using wireless sensor networks," J Health Med Inform, Vol. 4, No. 2, 2013. Available: http://www.omicsonline.org/a-hospital-healthcare-monitoring-system-using-wireless-sensor-networks-2157-7420.1000121.pdf

[2] W. Yao, C. H. Chu, and Z. Li, "The use of RFID in healthcare: Benefits and barriers," IEEE International Conference on RFID-Technology and Applications (RFID-TA), Guangzhou, China, June 17-19, 2010.

[3] M. Alam, M. Hafner, M. Memon, and P. Hung, "Modeling and enforcing advanced access control policies in healthcare systems with sectet," Workshop on Model-Based Trustworthy Health Information System (MOTHIS), 2007.

[4] Sun Microsystems, "Sun Java™ system RFID software 3.0 developer‘s guide," February 2006, [Online], Available: http://download.java.net/general/sun-rfid/Release30/Docs/Developers_Guide_819-4686.pdf

[5] Microsoft, "BizTalk server 2006 developer productivity study," January 2007, [Online], Available: ww.microsoft.com/biztalk/en/us/rd.aspx

[6] B. S. Prabhu, X. Su, H. Ramamurthy, C. Chu, and R. Gadh, "WinRFID – A middleware for the enablement of Radio Frequency Identification (RFID) based applications," Wireless Internet for the Mobile Enterprise Consortium (WINMEC), Los Angeles, December 2005. Available: http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.91.8928

[7] C. Floerkemeier, C. Roduner, and M. Lampe, "RFID application development with the Accada middleware platform", IEEE Systems Journal, Special Issue on RFID Technology, Vol. 1, No. 2, December 2007. Available: http://www.vs.inf.ethz.ch/publ/papers/floerkem-rfidap-2007.pdf

[8] J. Radhika, and S. Malarvizhi, "Middleware approaches for wireless sensor networks: an overview," International Journal of Computer Science Issues (IJCSI), 2012, Vol. 9, N° 3, pp: 224-229.

[9] K. Sohraby, D. Minoli, and T. Znati, "Middleware for wireless sensor networks", Wireless Sensor Networks Technology, Protocols, and Applications, John Wiley & Sons, 2007. Available : http://www.knovel.com/web/portal/browse/display?_EXT_KNOVEL_DISPLAY_bookid=4513&VerticalID=0

[10] S. Hadim, and N. Mohamed, "Middleware for wireless sensor networks: a survey," IEEE International Conference on Communication System Software and Middleware (COMSWARE), New Delhi, India, January 8-12, 2006.

[11] J. Singh, and J. Bacon, "Event-based data dissemination control in healthcare," Electronic Healthcare, 2009, vol. 0001, pp: 167—174.

[12] J. E. Bardram, and H. B. Christensen, "Middleware for pervasive healthcare - a white paper," Aarhus Denmark, 2001. Available: http://www.pervasivecomputing.dk/publications/files/wmmc2001.PDF

[13] ORACLE, "Oracle SOA Suite for Healthcare Integration," October 2013, [Online], Available: http://www.oracle.com/us/products/middleware/soa/soa-suite-for-healthcare-wp-2046692.pdf

[14] M. E. Ajana, M. Boulmalf, H. Harroud, and M. Elkoutbi (2011). RFID Middleware Design and Architecture, Designing and Deploying RFID Applications, Dr. Cristina Turcu (Ed.), ISBN: 978-953-307-265-4, InTech, DOI: 10.5772/16917. Available: http://www.intechopen.com/books/designing-and-deploying-rfid-applications/rfid-middleware-design-and-architecture

[15] M. E. Ajana, H. Harroud, M. Boulmalf, and H. Hamam, "A policy based event management middleware for implementing RFID applications," in Proceedings of the fifth IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), Marrakech, Morocco, October 2009.

[16] M. E. Ajana, H. Harroud, M. Boulmalf, and M. El Koutbi, "FlexRFID middleware in the supply chain: Strategic values and challenges,"

Contemporary Challenges and Solutions for Mobile and Multimedia Technologies. IGI Global, 2013. doi:10.4018/978-1-4666-2163-3.ch010.

[17] M. E. Ajana, H. Harroud, M. Boulmalf, M. Elkoutbi, A. Habbani, "Emerging wireless technologies in e-health trends, challenges, and framework design issues," Proceedings of International Conference on Multimedia Computing and Systems, International Conference on Multimedia Computing and Systems (ICMCS), Tangiers, Morocco, October 10-12, 2012.

[18] Q. Wang and T. Zhang, "Sec-SNMP: Policy-based security management for sensor networks", in Proceedings of the International Conference on Security and Cryptography (SECRYPT), 2008, pp. 222-226.

[19] M. Chraibi, H. Harroud, and A. Karmouch, "Personalized security in mobile environments using software policies," Proceedings of the 9th International Conference on Advances in Mobile Computing and Multimedia (MoMM), Hue City, Vietnam, December 5-7, 2011.

[20] M. Ahlsen, S. Asanin, P. Kool, P. Rosengren, and J. Thestrup, "Service-oriented middleware architecture for mobile personal health monitoring," Proceedings of the 2nd International ICST Conference on Wireless Mobile Communication and Healthcare (MOBIHEALTH 2011), Kos, Greece, October 5-7, 2011.

[21] Y. Y. Ou, P. Y. Shih, Y. H. Chin, T. W. Kuan, J. F. Wang, and S. H. Shih, "Framework of ubiquitous healthcare system based on cloud computing for elderly living," Proceedings of International Conference on Signal and Information Processing Association Annual Summit and Conference (APSIPA), Kaohsiung, Asia-Pacific, October 29 - November 1, 2013.

[22] V. Raychoudhurya, et al., "Middleware for pervasive computing: a survey," Pervasive and Mobile Computing, 2013, Vol. 9, No. 2, pp. 177–200.

[23] C. Park, et al., "RFID middleware evaluation toolkit based on a virtual reader emulator", in proceedings of the 1st International Conference on Emerging Databases, Busan, Korea, August 2009.

[24] M. E. Ajana, et al., "FlexRFID: A flexible middleware for RFID applications development," in proceedings of the 6th International Wireless and Optical Networks Communications (WOCN) Conference, Cairo, Egypt, April 2009.

[25] G. Schiele, et al., "Pervasive computing middleware", *Handbook of Ambient Intelligence and Smart Environments (AISE)*, Springer, US, 2010, pp. 201-227.

Modelling and Simulation of a Biometric Identity-Based Cryptography

Dania Aljeaid
School of Science and Technology
Nottingham Trent University
Nottingham, United Kingdom

Xiaoqi Ma
School of Science and Technology
Nottingham Trent University
Nottingham, United Kingdom

Caroline Langensiepen
School of Science and Technology
Nottingham Trent University
Nottingham, United Kingdom

***Abstract*—Government information is a vital asset that must be kept in a trusted environment and efficiently managed by authorised parties. Even though e-Government provides a number of advantages, it also introduces a range of new security risks. Sharing confidential and top-secret information in a secure manner among government sectors tends to be the main element that government agencies look for. Thus, developing an effective methodology is essential and it is a key factor for e-Government success. The proposed e-Government scheme in this paper is a combination of identity-based encryption and biometric technology. This new scheme can effectively improve the security in authentication systems, which provides a reliable identity with a high degree of assurance. This paper also demonstrates the feasibility of using finite-state machines as a formal method to analyse the proposed protocols. Finally we showed how Petri Nets could be used to simulate the communication patterns between the server and client as well as to validate the protocol functionality.**

Keywords—e-Government; identity-based cryptosystem; biometrics; mutual authentication; finite-state machine; Petri net.

I. Introduction

e-Government mainly acts as a communication bridge, whether from government to citizen, government to government, or government to business, in efficient and reliable ways through effective use of information technology. The main challenge in e-government is to develop a framework which promotes exchanging data securely among government agencies. While e-government provides a number of advantages, it also introduces a range of new security risks. Sharing confidential and top-secret information in a secure manner among government sectors tend to be the main element that government agencies look for.

When e-Government systems were being designed, **P**ublic **K**ey **I**nfrastructure (PKI) seemed to be the best solution for the scheme. PKI is presently deployed in most e-Government implementations, as it is perceived as a mature technology, which is widely supported and can be easily integrated with different systems. Examples of e-government initiatives that apply PKI on a large scale are the US eGov initiative (www.usa.gov) supported by Federal PKI [1] and the Saudi Arabian e-Government Program (yesser.gov.sa) [2].

One of the main issues concerning the security perspective in e-Government is to grant access to authorised users as well as the need to verify that the user is really who they claim to be. The most common solution to this problem is to deploy a PKI [3] and digital signatures in large-scale e-Government systems. Even though PKI supports strong authentication and digital signature, it has a few disadvantages. For example, users must be pre-enrolled, certificate directories can leak some critical information, key recovery is difficult and costly and boundary services (anti-spam, anti-virus, archiving) integration is very difficult [4].

Thus, to take full advantage of the capabilities of e-Government, end users need robust security solutions to achieve assurance when dealing with e-Government systems. A variant of public key cryptography that derives public keys directly from unique identity information (such as an e-mail address) known by the user is called **I**dentity-**B**ased **C**ryptography (IBC). This approach has recently received considerable attention from researchers [5, 6, 7, 8, 9], as the development of ID-Based Cryptography offers great flexibility and obviates the requirement for user certificates, since the identity of the user can be transformed into encryption keys and used for authentication.

To develop a new secure cryptosystem for e-Government, several schemes were investigated to determine which protocol would be suitable for the research. We propose a biometric-ID-based scheme using **E**lliptic **C**urve **C**ryptosystem (ECC), which is an improved combination scheme derived from two schemes [10, 11]. The proposed scheme is secure under the **C**omputational **D**iffie–**H**ellman **A**ssumption (CDHA) and tackles the security drawbacks of He *et al*.'s scheme and Li and Hwang's scheme. To overcome these, we applied a symmetric key cryptosystem to prevent attackers from altering or gaining any important information in the login and authentication messages.

The structure of this paper is organised as follows. In Section 2, we review related works on ID-Based Cryptography and Biometric authentication and briefly describe both He *et al*.'s and Li and Hwang's schemes. In Section 3, we design the new Biometric-ID-based Authentication Scheme. In Section 4, we model the new protocol with finite-state machines. In Section 5, we model the new protocol with Petri Nets to simulate the communication. We then provide a brief discussion on security analysis and comparisons with related schemes in Section 6. Finally, the conclusion is given in Section 7.

II. Review of Related Work

Without a secure and trusted infrastructure, organisations such as governments would leave data electronically unsecured and vulnerable to attacks. Therefore, governments are constantly looking for ways to deliver secure and reliable services. ID-Based Cryptography introduces a lightweight key management and offers encryption for data confidentiality and robust authentication, which are prerequisites for securing high-value transactions.

The idea of ID-based cryptography was originally proposed by Shamir in 1984 [12], but practical ID-based encryption schemes were not developed until recently. In 2001, Boneh & Franklin [5] developed a fully functional ID-based encryption scheme which can be constructed efficiently by using Weil pairing on elliptic curves. In ID-based cryptosystems, there is a trusted third party called a **P**rivate **K**ey **G**enerator (PKG) who is responsible for generating the secret keys for all users. As a result, a PKG holds the users' private keys. If a PKG is malicious, it can impersonate any user and therefore decrypt any cipher text or forge signature on any message. This can lead to a problem known as key escrow [13, 14].

There is no question that the **I**dentity-**B**ased **E**ncryption (IBE) scheme brings many advantages such as eliminating the need to distribute public keys. The enforcement of the private key generation by the Private Key Generator raises concerns of key escrow and/or privacy surrounding the management of private keys. To address this particular problem of key escrow, an implementation of biometric identification systems can be used as a private key. Biometric technology and verification systems offer a number of benefits to government sectors and users [15, 16].

He *et al*. [10] proposed an ID-based remote mutual authentication with key agreement scheme on ECC. This protocol attempts to cope with many of the well-known security and efficiency problems. However, the scheme has a potential flaw that may lead to man-in-the-middle attack and impersonation attack [17, 18]. It can be seen that, if an attacker E eavesdrops and listens to the communication between S_i and C_i, then E can intercept a valid login request $M_1=\{ID_{C_i}, T_{C_i}$, M, $MAC_k(ID_{C_i}, T_{C_i} ,M)\}$ or $h(ID_i \parallel Xs)$ and masquerade as a legal user.

Biometric technologies are becoming fundamental elements in ensuring highly secure identification and personal verification solutions [15]. Biometric keys can be extracted from keystroke patterns, the human voice [19], fingerprints [20, 21], handwritten signatures [22], and facial characteristics [23].

Li and Hwang [11] proposed an efficient biometrics-based remote user authentication scheme using smart cards. The security of their scheme is based on one-way hash functions, biometric verification, smart card and it uses a nonce. The scheme is very efficient in computation cost, which has been proved to be relatively low compared with other related schemes [24, 25, 26, 27]. The scheme is composed of four phases: the registration phase, the login phase, the authentication phase and password change phase.

One of the key characteristics of the cryptographic hash function is that the outputs are very sensitive to small perturbations in their inputs. Hash functions cannot be applied directly when the input data are noisy such as biometrics [28]. Therefore, a secure one-way hash function cannot be used for biometric verification. In the login phase of Li-Hwang's scheme, the user computes $h(B_i)$ based on a personal biometric template B_i. Then the biometric authentication process relies on comparing the hash value $h(B_i)$ with f_i . However, the scheme does not seem to be able to handle natural variation in the biometrics. For example, when the user logs in, his fresh biometric sample has to match exactly the template recorded during the registration phase, which never happens in practice. Thus, the protocol is fundamentally flawed and does not fulfil the basic objectives of a biometric authentication protocol. As a result, this may prevent a legal user from passing biometric verification at the login phase. So, Li-Hwang's scheme is vulnerable to denial-of-service attack. The scheme is also prone to man-in-the-middle attack and impersonation attack. The attacker can cheat the server by impersonating the user or can impersonate the server to cheat the user without knowing any secret information [29, 30, 31]

Combining ID based cryptography with biometric techniques can effectively improve the security in authentication systems, which provides a reliable identity with a high degree of assurance. The biometric technology is regarded as a powerful solution due to its unique link to an individual identity, which almost impossible to fake. Thus, a biometric identity is an inherent trait, which will always remain with the person all the time. In another words, using biometric techniques in IBE will mean that the person will always have their private key available.

III. Proposed Scheme

This research will focus on secure e-Government systems and improve their authentication and communication. To guarantee the security of these distributed systems, biometrics verification and ID-based cryptography are used. The proposed protocol is based on the following assumptions:

- We assume that shared secrets in registration phase will never be disclosed.
- We assume that cryptographic algorithms are secure. For example, it is impossible to decrypt a ciphertext without prior knowledge of the secret key.
- We assume that both client and server are able to generate a random number securely.

The security of the proposed scheme is based on the intractability of the following two mathematical problems on elliptic curves [5, 10]:

(i) **C**omputational **D**iffie–Hellman **A**ssumption (CDHA): Given P, xP, $yP \in G$, it is hard to compute $xyP \in G$.

(ii) **C**ollision **A**ttack **A**ssumption 1 (k-CAA1): For an integer k, and $x \in Z_n*$, $P \in G$,given

$(P, xP, h_0, (h_1, (h_1+x)^{-1}P),\ldots,(h_k, (h_k+x^{)-1}P))$, where $h_i \in Z_n^*$, and distinct for $0 \le i \le k$, it is hard to compute $(h_0+x)^{-1}P$.

The proposed scheme consists of four phases: system initialising phase, registration phase, login phase, and authentication phase. The notations used throughout this paper are summarised in Table 1.

TABLE.I. NOTATIONS USED IN THIS PAPER

Symbol	Definition
C_i	User/Client /Computer
S_i	Server
R_i	Registration Centre
ID_{S_i}	Identity of Server
ID_{C_i}	Identity of user C
PW_{C_i}	User's password
Bio_{C_i}	Biometric template of C
Pub_K	Public Key
Pr_K	Private Key
$\|\|$	Message concatenation operation
p, n	Two large prime numbers
F_p	A finite field
E	An elliptic curve over a finite field F
G	The group of elliptic curve points on E
P	A point on elliptic curve E with order n
xP	Denotes point multiplication on elliptic curve
y	A piece of secret information maintained by the server
(x, Pub_K$_s$)	The server S's Private/Public key pair, where Pub_K$_s$ = xP
r_{C_i}, r_{S_i}	A random number chosen by the C_i and S_i respectively
$H(.)$	A secure one-way hash function
MAC$_k$(m)	The secure message authentication code of m under the key k
$\oplus$	XOR operation

A. System initializing phase

In this phase, we follow the steps in He *et al*.'s scheme where the server S_i generates parameters of the system.

Step 1: S_i chooses an elliptic curve equation $E_P(a, b)$.

Step 2: S_i selects a base point P with the order n over $E_P(a, b)$

Step 3: S_i selects its master key x and secret information y and computes public key Pub_K$_s$ = xP

Step 4: The server chooses four secure one-way hash functions $H_1(.)$, $H_2(.)$, $H_3(.)$, $H_4(.)$, where $H(.)$ is a known hash function that takes a string and assigns it to a point on the elliptic curve, i.e. $H(A)$ = QA on E, where C is usually based on the identity

- $H_1(.)$: a secure one-way hash function, where H_1: $\{0, 1\}^* \rightarrow Z^*_n$
- $H_2(.)$: a secure one-way hash function, where H_2: $\{0, 1\}^* \rightarrow Z^*_p$
- $H_3(.)$: a secure one-way hash function, where H_3: $\{0, 1\}^* \rightarrow Z^*_p$
- $H_4(.)$:a secure one-way hash function, where H_4: $\{0, 1\}^* \rightarrow Z^*_p$

The server also chooses a message authentication code MAC$_k$(m). Then, it keeps x private and publishes $\{F_p, E, n, P,$ Pub_K$_s$, $H_1, H_2, H_3, H_4,$ MAC$_k(m)\}$.

B. Registration Phase

A user C_i with identifier ID_{C_i} should be registered first before using the services provided by R_i. Users may use their employee number as an identity when contacting R_i for authorisation. In this phase, C_i needs to perform the following steps.

Step 1: User C_i inputs their ID_{C_i}, personal biometrics Bio_{C_i}, on a specific biometric device, and provides the password PW_{C_i} to R_i via a secure channel (or to the registration centre in person).

Step 2: R_i reads current timestamp T_{S_i}, and computes the following:

$$f_i = H_4(Bio_{C_i})$$
$$z_i = H_4(PW_{C_i} \,\|\, f_i)$$
$$e_i = H_4(ID_{C_i} \,\|\, y) \oplus z_i$$

Step 3: R_i computes C_i's private key using the system private key x and C_i's public key.

Pr_K$_{C_i}$ = $(x+ H_4\,(ID_{C_i}))^{-1}\,P \in G$

Pub_K$_{C_i}$ = $H_4\,((ID_{C_i}) + \mathrm{x})\,P = H_4\,((ID_{C_i})P +$ Pub_K$_s$)

Step 4: R_i stores $\{ID_{C_i}$, H_4 (.), Enc{ }$_a$/Dec{ }$_a$, f_i, e_i, τ, Pr_K$_{C_i}\}$ on a secure database and sends it to the user via a secure channel, where Enc{ }$_a$/Dec{ }$_a$ is a symmetric encryption with secret key a and and τ is a predetermined threshold [28] for biometric verification.

C. Login Phase

The user C_i sends a login request to the server S$_i$ and performs the following steps:

Step 1: C_i enters the ID_{C_i} and PW_{C_i}, and then S_i verifies the authenticity of client's identity and password.

Step 2: C_i submits the Bio_{C_i} on specific biometric device, and then verifies the following:

$$\begin{cases} \text{Accept if } d(Bio_{C_i}, Bio^*_{C_i}) < \tau \\ \text{Reject if } d(Bio_{C_i}, Bio^*_{C_i}) \ge \tau \end{cases}$$

Step 3: if the above does not hold, it means the biometric information does not match the template stored in the system. Thus Ci does not pass the biometric verification process and the authentication scheme is terminated. Otherwise, Ci passes the biometric verification and computes the following:

$f_i = H_4\,(Bio_{C_i})$

$z^{`}_{i} = H_4\,(PW_{C_i} \,\|\, f_i)$

$M_1 = e_i \oplus z^{`}_{i} = H_4\,(ID_{C_i} \,\|\, y)$

$W_1 = r_{C_i} . P$

$M_2 = r_{C_i} . \mathrm{Pr\_K}_{C_i}$

$M_3 = M_1 \oplus r_{C_i}$

Where $r_{C_i} \in Z^{*}{}_{n}$ is a random number generated by the user. For this step, the random value r_{C_i} is introduced to mask the hash of the secret value $\mathrm{H}_4(ID_{C_i} \,\|\, y)$.

Step 4: C_i computes $k = H_2\,(ID_{C_i}, T_{C_i}, W_1, M_2)$, where T_{C_i} is a timestamp denoting the current time.

Step 5: Finally, C_i encrypts the message $\{ID_{C_i}, T_{C_i}, W_1, M_3, \mathrm{MAC}_k(ID_{C_i}, T_{C_i}, W_1, M_3)\}_a$ and sends it to the server $S_{i.}$

D. Authentication Phase

After receiving the request login message, S_i and C_i will perform the following steps for mutual authentication.

Step 1: Si decrypts the message {IDCi, TCi, W1, M3, MACk(IDCi, TCi, W1, M3)}a, then checks the validity of IDCi and the freshness of TCi. The freshness of TCi is checked by performing T` – TCi ≤ ΔT, where T` is the time when Si receives the above message and ΔT is a valid time interval. The case where IDCi is not valid or TCi is not fresh, then Si aborts the current session.

Step 2: If Step 1 holds, Si computes the following:

$M_2 = (x + H_1(ID_{C_i})^{-1}\ W_1$

$= \mathrm{Pr\_K}_{C_i} . r_{C_i}$

$k = H_2\,(ID_{C_i}, T_{C_i}, W_1, M_2)$

S_i checks the integrity of $MAC_k(ID_{C_i}, T_{C_i}, W_1, M_3)$ with the key k. S_i will quit the current session if the check produces a negative result.

Step 3: If Step 2 holds, Si chooses a random number RSi ∈ Z*n and computes the following:

$M_4 = H_4\,(ID_{C_i} \,\|\, y)$

$W_2 = r_{S_i} . P$

$K_{S_i} = r_{S_i} . W_1$

The session key $sk = H_3\,(ID_{C_i}, T_{C_i}, T_{S_i}, W_1, W_2, K_{S_i})$, where T_{S_i} is a timestamp denoting the current time

$M_5 = M_3 \oplus M_4 = r_{C_i}$

$M_6 = M_4 \oplus r_{S_i}$

$M_7 = H_4(M_3 \,\|\, M_5)$

Where M_5 is the random value r_{C_i} of the user C_i and only S_i can unmask the value because it can compute $H_4\,(ID_{C_i} \,\|\, y)$

Step 4: Then, S_i encrypts the message $\{ID_{C_i}, T_{S_i}, W_2, M_6, M_7, MAC_k(ID_{C_i}, T_{S_i}, W_2, M_6, M_7)\}_a$ and sends it to C_i

Step 5: Upon receiving the S_i's message, C_i first decrypts $\{ID_{C_i}, T_{S_i}, W_2, M_6, M_7, MAC_k(ID_{C_i}, T_{S_i}, W_2, M_6, M_7)\}_a$, and checks the freshness of T_{S_i} is by performing $T' – T_{S_i} \leq \Delta T$, where $T^{`}$ is the time when C_i receives the above message and ΔT is the expected time interval for the transmission delay.

Step 6: C_i verifies whether $M_7 \stackrel{?}{=} H_4\,(M_3 \,\|\, r_{C_i})$ and checks the integrity of $MAC_k(ID_{C_i}, T_{S_i}, W_2, M_6, M_7)$ with the key k. C_i will quit the current session if the check produces a negative result.

Step 7: If it holds, C_i believes that S_i is authenticated and then computes the following:

$K_{C_i} = r_{C_i} . W_2$

The session key $sk = H_3(ID_{C_i}, T_{C_i}, T_{S_i}, W_1, W_2, K_{C_i})$

$M_8 = M_6 \oplus M_1 = r_{S_i}$

$M_9 = H_4(M_6 \,\|\, M_8)$

Where M_9 is the random value r_{S_i} of the server S_i and only the client C_i, which know $M_1 = H_4\,(ID_{C_i} \| \, y)$, can send back the correct hashed value of $M_9 = H_4\,(H_4\,(ID_{C_i} \| \, y) \oplus r_{S_i}) \,\|\, r_{S_i})$

Step 8: C_i sends the encrypted message $\{M_{9,}\ MAC_k(M_9)\ \}_a$ to S_i

Step 9: After receiving C_i's message, S_i decrypts $\mathrm{Enc}\{M_9\}_a$ and check the integrity of $MAC_k(M_9)$. Then, S_i verifies whether $M_9 \stackrel{?}{=} H_4\,(M_6 \,\|\, r_{S_i})$

Step 10: If the above mentioned holds, S_i accept C_i's login request or otherwise rejects it

IV. Behaviour Modelling and State Machine

Verification is a crucial step in designing security protocols. A **F**inite-**S**tate **M**achine (FSM) is a powerful tool to simulate software architecture and communication protocols. FSM can only model the control part of a system and consists of a finite number of states, finite number of events, and finite number of transitions. An FSM may be regarded as a five-tuple [32]: $(Q, \sum, \Delta, \sigma, q_0)$, where:

- Q: finite set of symbols denoting states
- $\sum$: set of symbols denoting the possible inputs
- Δ: set of symbols denoting the possible outputs
- σ: transition function mapping to Qx$\sum$ to QxΔ
- $q_0 \in Q$: initial state.

The FSM is used to model the communication channel of proposed protocol between the Client C_i and the Server S_i.

Since the exchange of packets follows a pattern defined by a finite set of rules, it will be described by creating three finite-state machines $\text{FSM}_{\text{server}}$, $\text{FSM}_{\text{register}}$ and $\text{FSM}_{\text{client}}$.

A. Server FSM

The FSM at the server side represents the various on-going communications with the client at any point of time. It is modelled using 10 states and 22 transitions as detailed below. Fig. 1 shows the transitions diagram for the $\text{FSM}_{\text{server}}$.

1) The FSM_{server} will loop itself as the server is waiting for clients. The machine advances to the next state once it is triggered by a login/enrol transition accordingly.

2) When the FSM_{server} is in the state S1, it checks the validity of the received ID. If ID proved to be incorrect, S_i will request C_i to enter the valid ID for three times and FSM_{server} will loop until C_i enters the valid ID or if the attempts exceed three times. In the latter case, the C_i's account will be blocked and FSM_{server} changes state to S4 from state S1. Generally, three attempts are made through our protocol steps to allow common errors.

3) When the FSM_{server} is in the state S2, it is triggered by valid ID and it is now waiting for a valid PW. Once S_i receives PW, it verifies its validity. If PW proved to be wrong, S_i will request C_i to enter the valid PW for three times and FSM_{server} will loop until C_i enters the valid PW or if the attempts exceed three times. In the latter case, the C_i's account will be blocked and FSM_{server} changes state to S4 from state S2.

4) When the FSM_{server} is in the state S3, it is triggered by valid PW and it is now waiting for a valid Bio. Once Si receives Bio, it verifies its validity by comparing the imprinted Bio with the template stored. If Bio does not match the stored template, S_i will request C_i to enter the valid Bio up to three times and the FSM_{server} will loop until C_i enters the valid PW or if the attempts exceed three times. In the latter case, the C_i's account will be blocked and the FSM_{server} changes state to S4 from state S3.

5) In state S5, the FSM_{server} waits until receiving the login request SYN = $\{ID_{C_i}, T_{C_i}, W1, M3, MAC_k(ID_{C_i}, T_{C_i}, W_1, M_3)\}_a$ from the FSMclient to establish a connection by performing three-ways-handshake.

6) While in State S5, the FSM_{server} checks the validity of ID, freshness of T and the integrity of MAC_k. Then S_i generates a random number and timestamp in order to calculate the session key sk = $H_3(ID_{C_i}, T_{C_i}, T_{S_i}, W_1, W_2, K_{S_i})$. After that, Si replies SYN/ACK = $\{ID_{C_i}, T_{S_i}, W_2, M_6, M_7, MAC_k(ID_{C_i}, T_{S_i}, W_2, M_6, M_7)\}_a$ to the FSM_{client}.

7) In state S6, FSM_{server} waits until receiving ACK from the FSM_{client}. Once the FSM_{client} sends ACK = $\{M_9\}_a$, FSMserver verifies $M_9 \overset{?}{=} H_4\ (M_6 \mid\mid r_{S_i})$. At this instance, S_i authenticates C_i as a legitimate user.

8) At state S5 and state S6, FSM_{server} terminates the current session if any of the following situations occurs:

- The client ID is invalid
- The freshness of $\grave{T} - T_{C_i} \geq \Delta T$
- Negative result when checking the integrity of $MAC_k(ID_{C_i}, T_{C_i}, W_1, M_3)$
- If M_9 != $H_4\ (M_6 \parallel r_{S_i})$

At any stage of $\text{FSM}_{\text{server}}$, $\text{FSM}_{\text{server}}$ aborts the current session and changes to state S9 if the timeout exceeds the defined TIME_WAIT while waiting for packets. This feature helps to prevent an infinite wait when the $\text{FSM}_{\text{client}}$ fails to response.

B. Client FSM

The FSM at the client side represents the various on-going transmissions with the server at any point of time. It is modelled using 9 states and 21 transitions as detailed below. Fig. 1 shows the transitions diagram for the $\text{FSM}_{\text{client}}$.

1) First, the FSM_{client} is in the initial state C0 that is when the request for register/login is initiated by itself. While in state C0, the FSM_{server} checks whether C_i is enrolled or not. The next state will

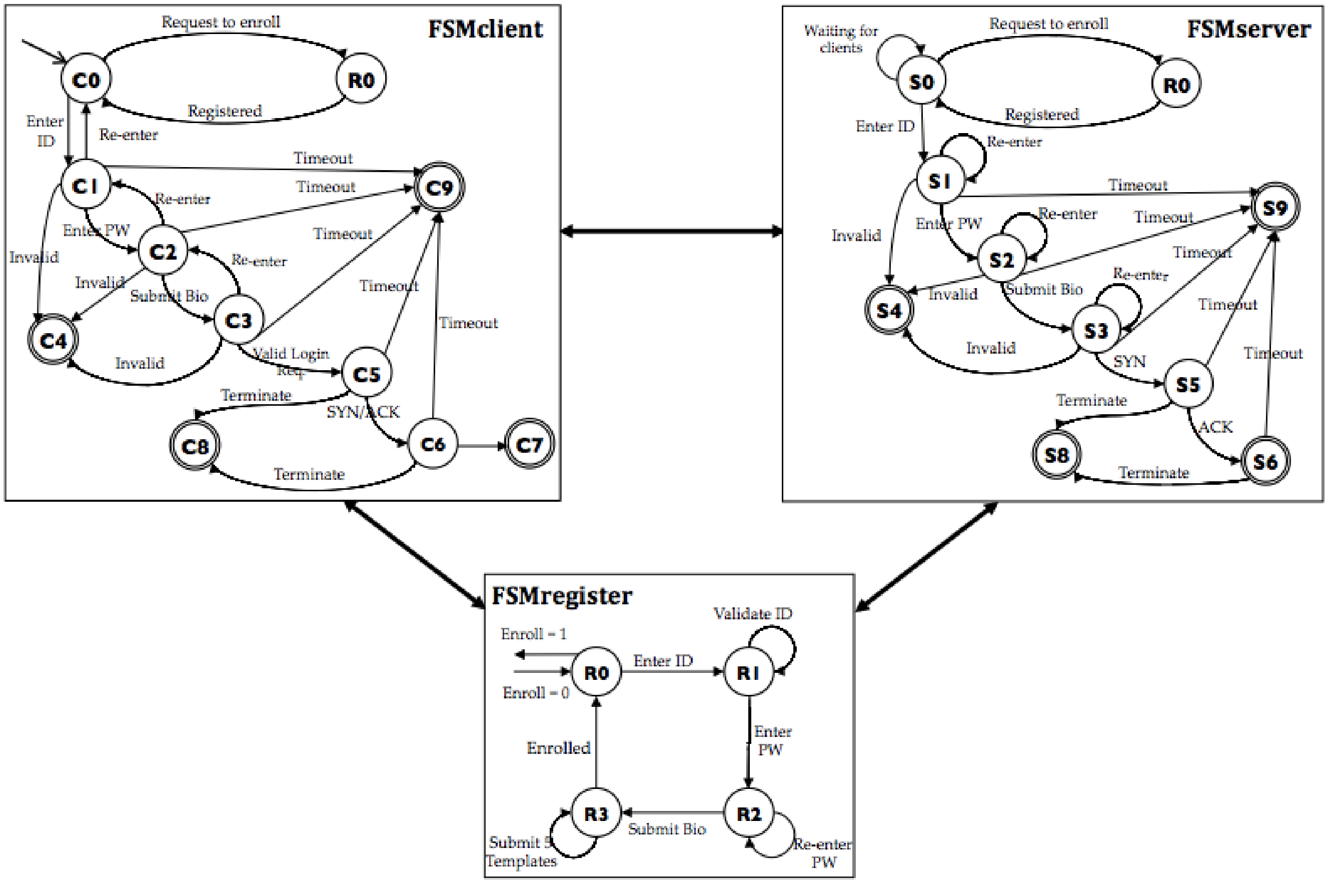


Fig.1. Proposed protocol FSM model

be decided according to the condition ClientReg == True.

2) In states C1, C2, C3, the FSM$_{client}$ is waiting for validating ID, PW, and Bio. Once the client credentials are validated, the FSM$_{client}$ triggers itself and changes to state C5.

3) In states C1, C2, C3, the client may require to re-enter ID, PW, Bio in case if they were incorrect. However, the client's account will be blocked if the number of attempts exceeds three trials, which change the above states to state C4.

- *ID*_attempt < 3, *ID*_attempt = *ID*_attempt +1
- *PW*_attempt < 3, *PW*_attempt = *PW*_attempt +1
- *Bio*_Attempt < 3, *Bio*_attempt = *Bio*_attempt +1

4) While in state C5, the FSM$_{client}$ is waiting for the FSM$_{server}$ response after sending the login request to establish the connection.

5) In state C6, the FSM$_{client}$ is validating the FSM$_{server}$ response by checking the integrity of MAC$_k$, ΔT and $M_7 \overset{?}{=} H_4 (M_4 || r_{C_i})$. If S_i is proved to be honest, C_i authenticates S_i at this stage.

6) While in state C6, the FSM$_{client}$ computes the shared session key sk = $H_3(ID_{C_i}, T_{C_i}, T_{S_i}, W_1, W_2, K_{C_i})$ and finalises the handshake procedure by sending ACK = $\{M_9\}_a$ to S_i.

7) In state C7, the FSM$_{client}$ is waiting to be authenticated by S_i.

8) In state C8, the client terminates the current session if one of the following occurs:

- Negative result when checking the integrity of *MAC$_k$*
- The freshness of $T^{`} – T_{S_i} \geq \Delta T$
- $M_7 \overset{?}{=} H_4 (M_4 || r_{C_i})$

At any stage of FSM$_{client}$, FSM$_{client}$ aborts the current session and changes to state *C*9 if the timeout exceeds the defined TIME_WAIT while waiting for packets. This feature helps to prevent an infinite wait when the FSM$_{server}$ fails to response.

C. Register FSM

The FSM at Registration side represents the various on-going transmissions with the server and client at any point of time. It is modelled using 4 states and 7 transitions as detailed below. Fig. 1 shows the transitions diagram for the FSM$_{register}$.

1) First, the FSM$_{register}$ is triggered if the client is not enrolled R0, that is when the request for register is initiated by

FSM_{client}. While in state C0, the FSM_{server} checks whether C_i is enrolled.

2) When once C_i enters ID, $FSM_{register}$ changes to state R1 and validates the format of ID. $FSM_{register}$ triggers itself. Then $FSM_{register}$ asks C_i to enter PW and changes to state R2.

3) In state R2, on receiving PW for the first time, $FSM_{register}$ requires C_i to re-enter PW for confirmation. Then it triggers and changes to the state R3.

4) In state R3, C_i is required to submit multiple scans of the biometric data to increase accuracy. Once the acquisition process is complete, $FSM_{register}$ trigger itself and sends a message to R0, which indicates that the enrolment is successful.

V. Protocol Model and Petri Nets

Due to the unique characteristics possessed by cryptographic protocols, analysis and evaluation tend to be more difficult than normal protocols. **P**etri **N**ets (PN) [33] offer a way to simulate the communication patterns between the server and client as well as to validate the protocol functionality.

Petri nets are a finite-state analysis approach that explicitly provides a graphical description for cryptographic protocols. The formal definition of a Petri net is shown in Table 2 [35]. Generally Petri nets focus on specific properties such as liveness, deadlock, livelock, boundedness and safeness [34,35,36]. Typically, a petri net must consist of the following components [35]:

- A set of *places* (drawn as circles in the graphical representation) represent conditions and possible states of the system.
- A set of *transitions* (drawn as rectangles or thick bars) represent a change of state which is caused by events or actions.
- A set of *arcs* (drawn as arrows) connecting a place to a transition and vice versa.
- *Tokens* (drawn as black dots) occupy places to represent the truth of the associated condition.

TABLE.II. Formal Definition of a Petri Net

A Petri net is 5-tuple, $PN=(P,T,F,W,M_0)$ where:

$P=\{p_1, p_2,\ldots,p_m\}$ is a finite set of places,
$T=\{t_1,t_2,\ldots,t_n\}$ is a finite set of transitions,
$F \subseteq (P \text{ X } T) \cup (T \text{ X } P)$ is a set of arcs (flow relations),
$W:F \rightarrow \{1, 2, 3,\ldots\}$ is a weight function,
$M_0:P \rightarrow \{0, 1, 2, 3,\ldots.\}$ is the initial marking,
$P \cap T= ø$ and $P \cup T \neq ø$.

A Petri net structure $N=(P, T, F, W)$ without any specific initial marking is denoted by N.

A Petri net with the given initial marking is denoted by (N,M_0).

Our technique involves simulation and verification by using Time-arc Petri nets. Initially, we build a PN model for client-server without intruder using TAPAAL simulation and verification software [37]. Moreover, it is worth to consider the following:

a) Define the places and transitions and declare their functionalities

b) Implement a token passing scheme once the initial marking is set.

c) Assess the model behaviour by examine reachability, boundedness, liveness.

The Petri net model in Fig. 2 represents the proposed protocol. The definitions of the places and transitions used in this model are illustrated in Table 3 and Table 4, respectively.

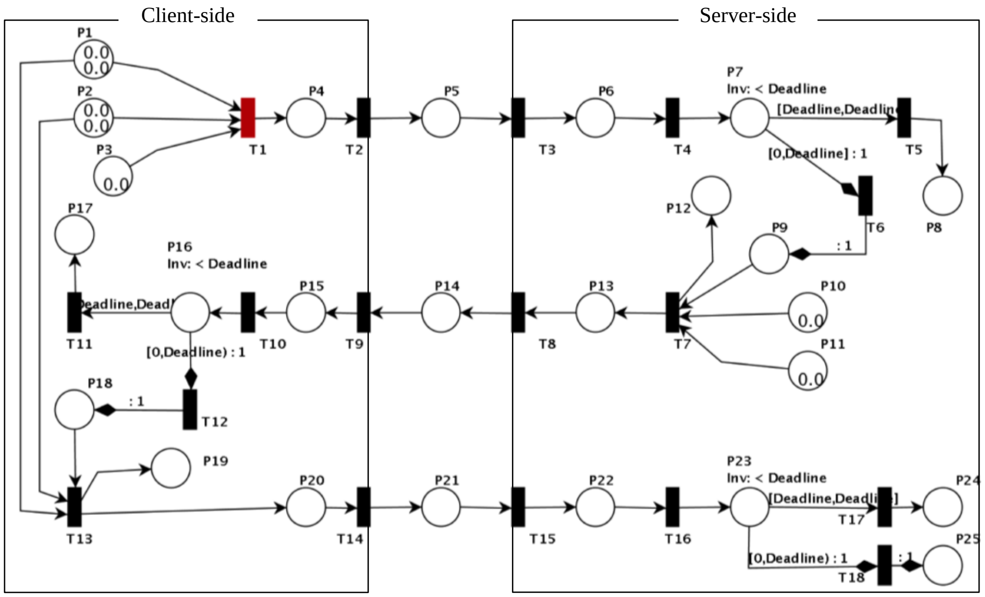
In our PN model, *places* mostly represent storage for requests, messages, ciphers, or session keys. *Transitions* represent actions that transform a current state to a new one. For example, the following events produce a new state: encryption, decryption, verification, and computations. *Tokens* are modelled in PN as shown in Fig. 2 to represent the key agreement and message exchange between the client and server. During simulation, the token firing rule imitates the three-way handshake procedure.

TABLE.III. Definitions of Places for The Proposed Model

Place	Definition	Place	Definition
P_1	Client random number	P_{14}	Encrypted SYN/ACK
P_2	Client timestamp	P_{15}	Decrypted SYN/ACK
P_3	SYN request	P_{16}	Verification message
P_4	Login request	P_{17}	Rejected request
P_5	Encrypted login request	P_{18}	Accept request – Server is authenticated
P_6	Decrypted login req.	P_{19}	Session key
P_7	Verification message	P_{20}	ACK
P_8	Rejected request	P_{21}	Encrypted ACK
P_9	Accepted request	P_{22}	Decrypted ACK
P_{10}	Server random number	P_{23}	Verification message
P_{11}	Server timestamp	P_{24}	Rejected request
P_{12}	Session Key	P_{25}	Accept request – Client is authenticated
P_{13}	SYN/ACK		

TABLE.IV. Definitions of Transitions for Proposed Model

Trans.	Definition	Trans.	Definition
T_1	Compute login request + SYN	T_{10}	Split the packet and verify
T_2	Encrypt	T_{11}	Drop the packet
T_3	Decrypt	T_{12}	Accept
T_4	Split the packet and verify	T_{13}	Compute ACK and session key
T_5	Drop the request	T_{14}	Encrypt ACK
T_6	Accept	T_{15}	Decrypt ACK
T_7	Compute SYN/ACK and session key	T_{16}	Split the packet and verify
T_8	Encrypt SYN/ACK	T_{17}	Drop the packet
T_9	Decrypt SYN/ACK	T_{18}	Accept



P={P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, P_9, P_{10}, P_{11}, P_{12}, P_{13}, P_{14}, P_{15}, P_{16}, P_{17}, P_{18}, P_{19}, P_{20}, P_{21}, P_{22}, P_{23}, P_{24}, P_{25}}

T = {T_1, T_2, T_3, T_4, T_5, T_6, T_7, T_8, T_9, T_{10}, T_{11}, T_{12}, T_{13}, T_{14}, T_{15}, T_{16}, T_{17}, T_{18}}

M_0: {2, 2, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,0}

Fig.2. The Petri net graph representing the new protocol.

After modeling the proposed protocol, it is essential to examine the behavioral properties of the model. Detailed behavioral properties for Petri nets can be found in [35]. Generating Reachability graph (Fig. 3) allows identifying the presence and absence behaviors of the modeled protocol.

A. *Reachability:*

Reachability or coverability can be conducted by numerating all states. In other words, deriving all the possible marking the protocol can reach in the model. This method can clearly identify all the enabled transition starting from the initial state and generating new states after firing transitions. The PN shown in Fig. 2 is bounded. This is evident from the reachability graph (Fig. 3), all set of reachable marking M_i, where i={0,1,2,…,19} are said to be reachable, that is to say there exists a sequence of transition firings which transform one marking state to another.

B. *Boundedness and safeness:*

Boundedness helps to detect overflows in the modeled system. This property is an indication of stability behavior of model. It is evident that the proposed PN is structurally bounded, for each place in the net hold at most 2 tokens given an initial marking m_0, that is to say that there are a finite number of states in the modeled protocol. Thus, the PN has no self-loop and satisfies the condition [35]:

A Petri net is k-bounded if all its places are k-bounded
A Petri net is structurally bounded if it is bounded in any initial marking

Hence, We can say that the PN is structurally 2-bounded, however, the PN is not safe because there are two nodes (P_1, P_2) contains more than one token. It does not fulfill the safeness condition, which is *1-boundedness.*

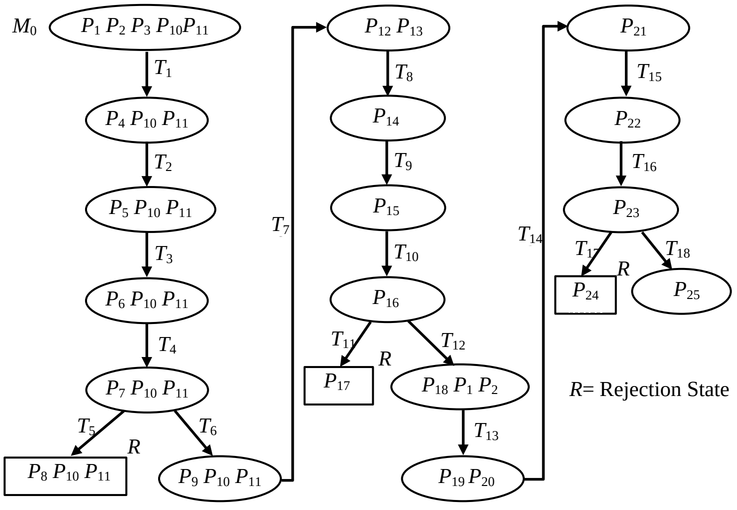


Fig.3. Reachability graph for proposed model

C. *Liveness*

The PN has a finite number of dead markings. The transitions (T_5, T_{11}, T_{17}) connected to places (P_8, P_{17}, P_{24}) respectively are not live if the protocol runs smoothly. Apart from that, the rest of places and their corresponding firing transitions are live. Occurrence of deadlocks (Rejection state) as shown in Fig. 3 is a result for aborting the current session between the client and serve; the token age exceeds the deadline.

Since PN contains deadlocks and not live, then the PN is considered not *reversible*.

VI. Security Analysis and Comparisions

The analysis suggests that the proposed scheme is well-designed for data confidentiality by using symmetric cryptography during the handshake procedure. Also, it ensures data integrity by applying a **M**essage **A**uthentication **C**ode function (MAC). Typically, the MAC function takes as input a secret key and data block and produces a hash value [38]. The client and server transmit the MAC value during the login and authentication phases. However, both client and server will be aware if an attacker alters the message because the integrity check of MAC value fails. When the communication session between C_i and S_i is over, the session key *sk* is discarded and a new session key is used in every protocol run to prevent a replay attack.

Mutual authentication and session key agreement

Based on FSM model and PN model, we proved that the protocol accomplished mutual authentication and secret session key agreement between a remote client and the server by establishing three-way challenge-response handshake technique. First, the client C_i sends the login request message $\{ID_{C_i}, T_{C_i}, W_1, M_3, MAC_k(ID_{C_i}, T_{C_i}, W_1, M_3)\}$ to the server S_i. Then S_i verifies the received message by checking the MAC integrity. After validating, S_i sends a challenge message $\{ID_{C_i}, T_{S_i}, W_2, M_6, M_{7,} MAC_k(ID_{C_i}, T_{S_i}, W_2, M_6, M_7)\}$ to C_i. Next, C_i check the validity of the received message $M_7 \overset{?}{=} H_4\,(M_4 \parallel r_{C_i})$ and accept or reject the server request according to the verification result. Finally, C_i sends a response message $M_9 = H_4(H_4(ID_{C_i} \parallel y) \oplus r_{S_i}) \parallel r_{S_i})$ to S_i. Upon receiving the message, S_i verifies if $M_9 \overset{?}{=} H_4\,(M_6 \parallel r_{S_i})$ holds. If so, S_i authenticates client C_i and allows him to get access. During the process, both S_i and C_i compute the session key $sk = H_3(ID_{C_i}, T_{C_i}, T_{S_i}, W_1, W_2, (r_{S_i}\,.r_{C_i}\,.P))$ successfully.

Denial-of-service-attack

Our scheme can withstand denial-of-service attack, because when the client C_i imprints personal biometrics $Bio^*_{C_i}$, the S_i will check the validity of $Bio^*_{C_i}$ with stored template by checking whether $d(Bio_{C_i}, Bio^*_{C_i}) < \tau$ holds. According to [31], the $Bio^*_{C_i}$ could pass the verification process even though there is some slight difference between $Bio_{C_i}, Bio^*_{C_i}$.

As for the computation cost, the proposed protocol is relatively low cost and efficient since only symmetric encryption; hash operations and XOR operations are required. Moreover, it is based on ECC which has significant advantages over other public-key cryptography. ECC provides the same security level of RSA cryptosystem but with a shorter key length and faster computation [39]

In Table 5, we summarised the performance and demonstrated comparisons between the proposed scheme and other related schemes. The evaluation parameters are defined in Table 6. Even though the number of operations is more than in other schemes, our scheme holds other security properties. The proposed protocol is based on a two-factor user authentication mechanism and it is obvious that it takes few more hash operations and XOR operations for the server and client. Due to the security weaknesses in related schemes, we applied symmetric encryption and symmetric decryption to ensure the confidentiality and the integrity of transmitted packets. Therefore this feature makes the proposed scheme effective.

TABLE.V. PERFORMANCE COMPARISONS

	He et al.'s Scheme	Li-Hwang's Scheme	Proposed Scheme
Client	$2T_H + 2T_{MAC}$	$3T_H + 3T_X$	$6T_H + 3T_X + 3T_{MAC} + 2T_{SE} + T_{SD}$
Server	$4T_H + 2T_{MAC}$	$4T_H + 2T_X$	$6T_H + 2T_X + 3T_{MAC} + 1T_{SE} + 2T_{SD}$

TABLE.VI. EVALUATION PARAMETERS

Symbol	Definition
T_X	Time for executing an XOR operation
T_H	Time for executing a one-way hash function
T_{MAC}	Time for executing a message authentication code
T_{SE}	Time for executing a symmetric encryption operation
T_{SD}	Time for executing a symmetric decryption operation

VII. CONCLUSION AND FUTURE WORK

The paper demonstrates how a combination of ID-based encryption with biometrics can be effective and more suited to e-Government environments. Moreover, the new biometric-identity-based scheme can be integrated into e-Government systems as the main authentication method and for secure communication as well. The proposed scheme is aimed to initiate secure authentication and communication between the client and server by building a robust mechanism between communicating government parties. The presented protocol is described as a three-way handshake procedure to establish a reliable connection and ensure secure data sharing. Moreover, we have simulated and validated the behaviour of the proposed protocol by using finite-state machines and Petri nets.

In future, an in-depth security analysis and evaluation will be conducted to thoroughly assess for security vulnerabilities and weaknesses. Furthermore, it is essential to consider using Petri Nets to add an intruder model and implement a token-passing scheme. At this stage, we will examine different attacks, such as impersonation attack, man-in-the-middle attack, and replay attack against the proposed scheme and verify its security.

ACKNOWLEDGMENT

This research has been funded by Saudi Arabian Cultural Bureau in London and King Abdul Aziz University in Saudi Arabia.

REFERENCES

[1] Caloyannides, M., authentication framework and programs. IT professional, , pp. 16-21.

[2] Sahraoui, S., Gharaibeh, G. And Al-Jboori, A., 2006. E-Government in Saudi Arabia: Can it overcome its challenges, e-Government Workshop 2006.

[3] Evans, D. And Yen, D.C., 2005. E-government: An analysis for implementation: Framework for understanding cultural and social impact. Government Information Quarterly, 22(3), pp. 354-373.

[4] Voltage Security, 2006. Identity-Based Encryption and PKI Making Security Work. http://www.voltage.com/pdf/IBE_and_PKI.pdf edn. Voltage Security, Inc.

[5] Boneh, D. And Franklin, M., 2001. Identity-based encryption from the Weil pairing, Advances in Cryptology—CRYPTO 2001 2001, Springer, pp. 213-229.

[6] Gentry, C. and Silverberg, A., 2002. Hierarchical ID-based cryptography. Advances in Cryptology—ASIACRYPT 2002, , pp. 149-155.

[7] Al-Riyami, S. and Paterson, K., 2003. Certificateless public key cryptography. Advances in Cryptology-ASIACRYPT 2003, , pp. 452-473.

[8] Boneh, D. and Boyen, X., 2004. Efficient selective-ID secure identity-based encryption without random oracles, Advances in Cryptology-EUROCRYPT 2004 2004, Springer, pp. 223-238.

[9] Duffy, A. And Dowling, T., 2004. An object oriented approach to an identity based encryption cryptosystem, Software Engineering and Applications 2004, ACTA Press.

[10] He, D., Chen, J. And Hu, J., 2012. An ID-based client authentication with key agreement protocol for mobile client–server environment on ECC with provable security.Information Fusion, 13(3), pp. 223-230.

[11] Li, C.T. And Hwang, M.S., 2010. An efficient biometrics-based remote user authentication scheme using smart cards. Journal of Network and Computer Applications, 33(1), pp. 1-5.

[12] Shamir, A., 1985. Identity-based cryptosystems and signature schemes, *Advances in cryptology* 1985, Springer, pp. 47-53.

[13] Liao, J., Xiao, J., Qi, Y., Huang, P. and Rong, M., 2005. ID-based signature scheme without trusted PKG, *Information Security and Cryptology* 2005, Springer, pp. 53-62.

[14] Yuen, T., Susilo, W. and Mu, Y., 2010. How to construct identity-based signatures without the key escrow problem. *Public Key Infrastructures, Services and Applications,* , pp. 286-301.

[15] Vacca, J.R., 2007. *Biometric technologies and verification systems.* Oxford: Butterworth-Heinemann.

[16] Vielhauer, C., 2005. Biometric user authentication for IT security: from fundamentals to handwriting. New York ; London: Springer.

[17] Islam, S.H. and Biswas, G., 2012. An improved ID-based client authentication with key agreement scheme on ECC for mobile client-server environments. *Theoretical and Applied Informatics,* **24**(4), pp. 293-312.

[18] Wang, D. and Ma, C., 2013. Cryptanalysis of a remote user authentication scheme for mobile client-server environment based on ECC. *Information Fusion,* .

[19] Monrose, F., Reiter, M.K., Li, Q. and Wetzel, S., 2001. Cryptographic key generation from voice, *Security and Privacy, 2001. S&P 2001. Proceedings. 2001 IEEE Symposium on* 2001, IEEE, pp. 202-213

[20] Seto, Y, 2002. Development of personal authentication systems using fingerprint with smart cards and digital signature technologies, The Seventh International Conference on Control, Automation, Robotics and Vision.

[21] Clancy T.C., Kiyavash, N. and D.J. Lin, D.J., 2003. *Secure smart card based fingerprint authentication*, Proceedings of the 2003 ACM SIGMM Workshop on Biometrics Methods and Application, WBMA.

[22] Hao, F, Chan, C.W., 2002. Private key generation from on-line handwritten signatures, Information Management & Computer Security, Issue 10, No. 2, pp. 159–164.

[23] Goh, A. and Ngo, D., 2003. Computation of cryptographic keys from face biometrics. *Communications and Multimedia Security.Advanced Techniques for Network and Data Protection,* , pp. 1-13.

[24] Hwang, M., Lee, C. and Tang, Y., 2002. A simple remote user authentication scheme. Mathematical and Computer Modelling, 36(1), pp. 103-107.

[25] Lin, C. and Lai, Y., 2004. A flexible biometrics remote user authentication scheme. Computer Standards & Interfaces, 27(1), pp. 19-23.

[26] Lee, N. and Chiu, Y., 2005. Improved remote authentication scheme with smart card. Computer Standards & Interfaces, 27(2), pp. 177-180.

[27] Chang, Y., Chang, C. and Su, Y., 2006. A secure improvement on the user-friendly remote authentication scheme with no time concurrency mechanism, Advanced Information Networking and Applications, 2006. AINA 2006. 20th International Conference on 2006, IEEE, pp. 5 pp.

[28] Inuma, M., Otsuka, A. and Imai, H., 2009. Theoretical framework for constructing matching algorithms in biometric authentication systems. *Advances in Biometrics.* Springer, pp. 806-815.

[29] Lu, J., Zhang, S. and QIE, S., 2011. Enhanced Biometrics-based Remote User Authentication Scheme Using Smart Cards. *IACR Cryptology ePrint Archive,* **2011**, pp. 676.

[30] Jeon, S., Kim, H. and Kim, M., 2011. Enhanced biometrics-based remote user authentication scheme using smart cards. *J.of Security Engineering,* **8**(2), pp. 237-254.

[31] Li, X., Niu, J., Ma, J., Wang, W. and Liu, C., 2011. Cryptanalysis and improvement of a biometrics-based remote user authentication scheme using smart cards. Journal of Network and Computer

[32] Hopcroft, J.E., Je rey D. Ullman. 1979. Introduction to automata theory, languages, and computation.

[33] Petri, C.A., 1962. Kommunikation mit Automaten. Ph. D. Thesis, University of Bonn.

[34] Peterson, J.L., 1981. Petri Net Theory and the Modeling of Systems. Prentice-Hall.

[35] Murata, T., 1989. Petri nets: properties, analysis and applications. Proceedings

[36] Bobbio, A., 1990. System modelling with Petri nets. *Systems reliability assessment.* Springer, pp. 103-143.

[37] TAPAAL 2.4.3 Petri nets simulation and verfication of timed-arc Petri nets. Available ar: www.tapaal.net.

[38] Stallings, W., 2011. Cryptography and Network Security, 5/E. Pearson Education, Inc.

[39] Yokoyama, V.T.V., 2000. Elliptic curve cryptosystem. Fujitsu Sci.Tech.J, 36(2), pp. 140-146.

A real time OCSVM Intrusion Detection module with low overhead for SCADA systems

Leandros A. Maglaras
Department of Computing, University of Surrey
Guilford, UK
Email: l.maglaras@surrey.ac.uk

Jianmin Jiang
Department of Computing, University of Surrey
Guilford, UK
Email: jianmin.jiang@surrey.ac.uk

***Abstract*—In this paper we present a intrusion detection module capable of detecting malicious network traffic in a SCADA (Supervisory Control and Data Acquisition) system. Malicious data in a SCADA system disrupt its correct functioning and tamper with its normal operation. OCSVM (One-Class Support Vector Machine) is an intrusion detection mechanism that does not need any labeled data for training or any information about the kind of anomaly is expecting for the detection process. This feature makes it ideal for processing SCADA environment data and automate SCADA performance monitoring. The OCSVM module developed is trained by network traces off line and detect anomalies in the system real time.**

In order to decrease the overhead induced by communicated alarms we propose a new detection mechanism that is based on the combination of OCSVM with a recursive k-means clustering procedure. The proposed intrusion detection module $\mathcal{K}-\mathcal{OCSVM}$ is capable to distinguish severe alarms from possible attacks regardless of the values of parameters σ and ν, making it ideal for real-time intrusion detection mechanisms for SCADA systems. The most severe alarms are then communicated with the use of IDMEF files to an IDSIDS (Intrusion Detection System) system that is developed under CockpitCI project. Alarm messages carry information about the source of the incident, the time of the intrusion and a classification of the alarm.

***Keywords*—SCADA systems; OCSVM; intrusion detection**

I. Introduction

Cyber-physical systems are becoming vital to modernizing the national critical infrastructure systems. Cyber attacks usually target valuable infrastructures assets, taking advantage of architectural/technical vulnerabilities or even weaknesses in the defense systems. While there is the case in some situations, most weaknesses in CIs arise from the fact that most CIs are adopting off-the-shelf technologies from the IT world, without a significant change in terms of the operator mindset, still based on the ”airgap” security principle that suggests that an apparently isolated and obscure system is implicitly secure. Once you open the system to off-the-shelf solutions, you also increase its exposure to cyber-attacks. Several techniques and algorithms have been reported by researchers for intrusion detection. One big family of intrusion detection algorithms is rule based algorithms. In real applications though, during abnormal situations, the behavior of the system cannot be predicted and does not follow any known pattern or rule. This characteristic makes rule based algorithms incapable of detecting the intrusion.

Generally, anomaly detection can be regarded as binary classification problem and thus many classification algorithms which are utilized for detecting anomalies, such as neural networks, support vector machines, K-nearest neighbour (KNN) and Hidden Markov model can be used. However, strictly speaking, they are not intrusion detection algorithms, as they require knowing what kind of anomaly is expecting, which deviates the fundamental object of intrusion detection. In addition these algorithms may be sensitive to noise in the training samples.

Segmentation and clustering algorithms seem to be better choices because they do not need to know the signatures of the series. The shortages of such algorithms are that they always need parameters to specify a proper number of segmentation or clusters and the detection procedure has to shift from one state to another state. Negative selection algorithms on the other hand,are designed for one-class classification; however, these algorithms can potentially fail with the increasing diversity of normal set and they are not meant to the problem with a small number of self-samples, or general classification problem where probability distribution plays a crucial role. Furthermore, negative selection only works for a standard sequence, which is not suitable for online detection. Other algorithms, such as time series analysis are also introduced to anomaly detections, and again, they may not be suitable for most of the real application cases.

To minimize the above mention drawbacks an intelligent approach based on OCSVM [One-Class Support Vector Machine] principles are proposed for intrusion detection. OCSVM is a natural extension of the support vector algorithm to the case of unlabeled data, especially for detection of outliers. The OCSVM algorithm maps input data into a high dimensional feature space (via a kernel) and iteratively finds the maximal margin hyperplane which best separates the training data from the origin (Figure 1).

OCSVM principles have shown great potential in the area of anomaly detection [1]–[4]. IDS can provide active detection and automated responses during intrusions [5]. Commercial IDS products such as NetRanger, RealSecure, and Omniguard Intruder alert work on attack signatures. These signatures needed to be updated by the vendors on a regular basis in order to protect from new types of attacks. Most of the current intrusion detection commercial softwares are based on approaches with statistics embedded feature processing, time series analysis and pattern recognition techniques. Several extensions of OCSVM method have been introduced lately [6]–[8]

A. Contributions

The present article develops a intrusion detection method, namely the *K-means OCSVM ($\mathcal{K}-\mathcal{OCSVM}$)*. Using the well known OCSVM method with default values for parameters σ and ν we distinguish real from false alarms with the use of a recursive k-means clustering method. This is very different from all previous methods that required pre-selection of parameters with the use of cross-validation or other methods that ensemble of One class classifiers [9].

The article makes the following contributions:

- OCSVM is tested for intrusion detection in SCADA system
- A new one class classifier $\mathcal{K}-\mathcal{OCSVM}$ is proposed.
- The proposed classifier combines OCSVM with RBF kernel with a recursive K-means clustering method.
- $\mathcal{K}-\mathcal{OCSVM}$ separates in real time false from real alarms.
- A performance evaluation of OCSVM and the proposed method with different parameters is conducted.

The rest of this article is organized as follows: Section II describes the OCSVM method. Section III describes the proposed $\mathcal{K}-\mathcal{OCSVM}$ method; In Section IV the use of OCSVM in SCADA systems is presented; Section V presents the features extracted from the network traces for the testing and training of the model Section VI describes how our model is integrated in the IDS system, section VII presents the evaluation of OCSVM and $\mathcal{K}-\mathcal{OCSVM}$ methods and results. Section VIII concludes the article.

II. OCSVM METHOD

The one-class classification problem is a special case of the conventional two-class classification problem, where only data from one specific class are available and well represented. This class is called the target class. Another class, which is called the outlier class, can be sampled very sparsely, or even not at all. This smaller class contains data that appear when the operation of the system varies from the normal, due to a possible attack. In general cases, the outlier class might be very difficult or expensive to measure. Therefore, in the one class classifier training process, mainly samples from the target class are used and there is no information about its counterpart. The boundary between the two classes has to be estimated from data in the only available target class. Thus, the task is to define a boundary around the target class, such that it encircles as many target examples as possible and minimizes the chance of accepting outliers.

Scholkopf et al. [10] developed an OCSVM algorithm to deal with the one-class classification problem. The OCSVM may be viewed as a regular two-class SVM, where all the training data lie in the first class, and the origin is taken as the only member of the second class. The OCSVM algorithm first maps input data into a high dimensional feature space via a kernel function and then iteratively finds the maximal margin hyperplane, which best separates the training data from the origin. Thus, the hyperplane (or linear decision boundary) corresponds to the classification function

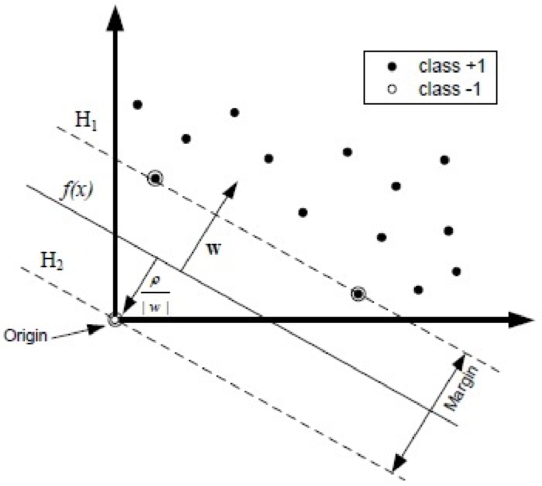


Fig. 1. OCSVM maps input data into a high dimensional feature space

$$f(x) =< x, w > +b \tag{1}$$

where w is the normal vector and b is a bias term. The OCSVM solves an optimization problem to find the function f with maximal geometric margin. This classification function can be used to assign a label to a test example x. If $f(x) < 0$, then x is labeled as an anomaly (outlier class), otherwise it is labeled normal (target class).

Using kernel functions, solving the OCSVM optimization problem is equivalent to solving the following dual quadratic programming problem.

$$min \frac{1}{2} \sum_{i,j} a_i a_j K(x_i, x_j) \tag{2}$$

subject to $0 \leq a_i \leq 1/\nu l$, and $\sum_i a_i \leq 1$.

Where a_i is a Lagrange multiplier, which can be thought of as a weight for example x, such that vectors associated with non-zero weights are called support vectors and solely determine the optimal hyperplane, ν is parameter that controls the trade-off between maximizing the number of data points contained by the hyperplane and the distance of the hyperplane from the origin, l is the number of points in the training dataset, and $K(x_i, x_j)$ is the kernel function.

Using the kernel function to project input vectors into a feature space, nonlinear decision boundaries are allowed. Generally, four types of kernel are often used: linear, polynomial, sigmoid and Gaussian radial basis function (RBF) kernels. In this paper, we use the RBF kernel, which has been commonly used for the OCSVM.

$$K(x_i, x_j) = exp(-\sigma ||x_i - x_j||^2), \quad \sigma > 0 \tag{3}$$

Although the OCSVM requires samples of the target class only as training samples, some studies showed that when negative examples (i.e. samples of outlier classes) are available, they can be used during the training to improve the performance of the OCSVM. In this paper only normal data were used for the training of the method, though a similar to the one proposed by Tax [11], which includes a small amount of samples of the outlier class, will be also applied and evaluated in the near future.

For the OCSVM with an RBF kernel, two parameters σ and ν need to be carefully selected in order to obtain the optimal classification result. A common strategy is to separate the data set into two parts, of which one is considered unknown. The prediction accuracy obtained from the unknown set more precisely reflects the performance on classifying an independent data set. An improved version of this procedure is known as cross-validation. Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.

In ν-fold cross-validation, the training set is divided into ν subsets of equal size. Sequentially one subset is tested using the classifier trained on the remaining ν .. 1 subsets. Thus, each instance of the whole training set is predicted once so the cross-validation accuracy is the percentage of data which are correctly classified. The cross-validation procedure can prevent the over fitting problem.

Unnthorsson et al. [12] proposed another method to select parameters for the OCSVM. In their method, ν was first set to a user-specified allowable fraction of misclassification of the target class (e.g. 1% or 5%), then the appropriate σ value was selected as the value for the classification accuracy curve of training samples first reaches $1 - \nu$. The obtained ν and σ combination can then be used in the OCSVM classification.

III. $\mathcal{K}-\mathcal{OCSVM}$

OCSVM similar to other one-class classifiers suffer from false positive and over fitting. The former is a situation that occurs when the classifier fires an alarm in the absence of real anomaly in the system and happens when parameter σ has too large vale. The latter is the situation when a model begins to memorize training data rather than learning to generalize from trend and it shows up when parameter σ is given relatively small value [13].

In this article we propose the combination of OCSVM method with a recursive k-means clustering, separating the real from false alarms in real time and with no pre-selection of parameters σ and ν. The proposed $\mathcal{K}-\mathcal{OCSVM}$ combines the well known OCSVM classifier with the RBF kernel with a recursive K-means clustering module. Figure 2 illustrates the procedure of intrusion detection of our proposed $\mathcal{K}-\mathcal{OCSVM}$ model.

The OCSVM classifier runs with default parameters and the outcome consists of all possible outliers. These outliers are clustered using the k-means clustering method with 2 clusters, where the initial means of the clusters are the maximum and the minimum negative values returned by the OCSVM module. From the two clusters that are created from the K-means clustering, the one that is closer to the maximum negative value (severe alerts) is used as input in the next call of the K-means clustering. This procedure is repeated until all outcomes are put in the same cluster or the divided set is big enough compared to the initial one, according to the threshold parameter k_{thres}.

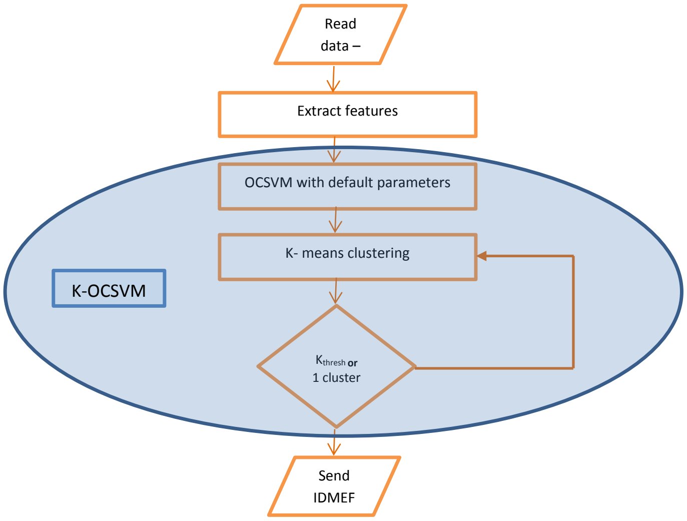


Fig. 2. $\mathcal{K}-\mathcal{OCSVM}$ module

K-means clustering method divides the outcomes according to their values and those outcomes with most negative values are kept. That way, after the completion of this recursive procedure only the most severe alerts are communicated from the $\mathcal{K}-\mathcal{OCSVM}$. The division of the data need no previous knowledge about the values of the outcomes which may vary from -0.1 to -160 depending of the assigned values to parameters σ and ν. The method can find the most important/possible outliers for any given values to parameters σ and ν.

One important parameter that affects the performance of $\mathcal{K}-\mathcal{OCSVM}$ is the value of threshold k_{thres}. For given value 2, the final cluster of severe alerts that the method communicates to other parts of the IDS system is limited. For bigger value (3 or more) the number of alerts rises till the method degrades to the initial OCSVM. The optimal value for the given parameter k_{thres} is a matter for future investigation.

IV. OCSVM FOR SCADA SYSTEM

Cyber-attacks against SCADA systems [14] are considered extremely dangerous for Critical Infrastructure (CI) operation and must be addressed in a specific way [15], [16]. Presently one of the most adopted attacks to a SCADA system is based on fake commands sent from the SCADA to the RTUs. OCSVM [17]–[19] possesses several advantages for processing SCADA environment data and automate SCADA performance monitoring, which can be highlighted as:

- In the case of SCADA performance monitoring, which patterns in data are normal or abnormal may not be obvious to operators. Since OCSVM does not require any signatures of data to build the detection model it is well suited for intrusion detection in SCADA environment.

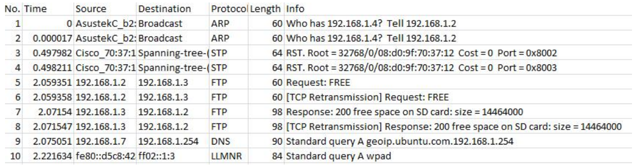
- Since the detection mechanism does not require any prior information of the expected attack types, OCSVM is capable of detection both known and unknown (novel) attacks.
- In practice training data, taken from SCADA environment, could include noise samples. Most of the classification based intrusion detection methods are very sensitive to noise. However, OCSVM detection approach is robust to noise samples in the training process.
- Algorithm configuration can be controlled by the user to regulate the percentage of anomalies expected.
- Due to the low computation time, OCSVM detectors can operate fast enough for online SCADA performance monitoring.
- Typically monitoring data of SCADA systems consists of several attributes and OCSVM is capable of handling multiple attributed data .

V. Attribute extraction

Feature extraction is essential in a classification problem. In order to train the OCSVM module properly we used a network trace file from a SCADA system. Based on the analysis of data that is presented in the upcoming subsections we selected some initial features that are used as attributes for our OCSVM model. We also found that additional features of the system must be combined with these initial ones in order to better represent the current state of the network operation.

A. Analysis of data

The dataset consists of about 1570 rows each one representing a send packet. There are several sources(12) and destinations(17). The packets use different protocols (13) each performing a different task in the network. Modbus/TCP protocol is a commonly available means of connecting industrial electronic devices (PLCs).



No.	Time	Source	Destination	Protocol	Length	Info
1	0	AsustekC_b2:	Broadcast	ARP	60	Who has 192.168.1.4? Tell 192.168.1.2
2	0.000017	AsustekC_b2:	Broadcast	ARP	60	Who has 192.168.1.4? Tell 192.168.1.2
3	0.497982	Cisco_70:37:1	Spanning-tree-(	STP	64	RST. Root = 32768/0/08:d0:9f:70:37:12 Cost = 0 Port = 0x8002
4	0.498211	Cisco_70:37:1	Spanning-tree-(	STP	64	RST. Root = 32768/0/08:d0:9f:70:37:12 Cost = 0 Port = 0x8003
5	2.059351	192.168.1.2	192.168.1.3	FTP	60	Request: FREE
6	2.059358	192.168.1.2	192.168.1.3	FTP	60	[TCP Retransmission] Request: FREE
7	2.07154	192.168.1.3	192.168.1.2	FTP	98	Response: 200 free space on SD card: size = 14464000
8	2.071547	192.168.1.3	192.168.1.2	FTP	98	[TCP Retransmission] Response: 200 free space on SD card: size = 14464000
9	2.075051	192.168.1.7	192.168.1.254	DNS	90	Standard query A geoip.ubuntu.com.192.168.1.254
10	2.221634	fe80::d5c8:42	ff02::1:3	LLMNR	84	Standard query A wpad

Fig. 3. Snapshot of SCADA dataset.

The addresses that create the biggest traffic (packets created) in the system are mainly two (192.168.1.2, 192.168.1.3) which represent probably PCs or PLCs. The distribution of sources and destinations is shown in Figure 4. The basic protocols used are: DNS, FTP, MDNS, MODBUS, TCP, UDP.

Figure 4 presents the source/destination distribution of the packets in the network. It is evident that most of the traffic is produced by only a small portion of the network (two or three sources) that send their packets to a limited set of nodes. This communication mainly constist of control packets exchanged between industrial based computers and Programmable Logic Controllers (PLCs). Thorough knowledge of the type of the source that produces the traffic and a reputation metric that represents the history of the source in terms of vulnerability would add additional intelligence to the detection module, making it capable of distinguishing between traffic bursts and actual attacks.

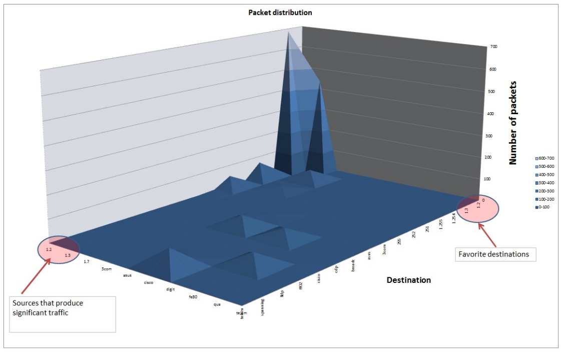


Fig. 4. Source - destination distribution

Protocols used for the communication among nodes of the network may also include information about a possible abnormal behavior of the system. Protocols are used for specific tasks in a network. Different attacks use different protocols. A proper filtering of the network data or use of protocol as an attribute for the OCSVM module, may be useful for the detection intrusion mechanism.

It is evident from Figure 5 that the accumulative packet size over time is almost a straight line with some sudden rises only is the three instances when big files are circulated in the system. These files are FTP files of size 590 bytes which is ten times larger than the usual send packets in the system. This smooth packet growth over time represents a normal behavior of the system when no malicious data is detected and the slop of the line can vary from time to time, depending on the load of the system. When malicious data are broadcasted over the network the accumulative packet size may rise, but since in many situations intruders send 0 lenght packet sizes (ack/nack packets) this feature by itself is not enough to detect these situations.

A burst in traffic injected in the system is another characteristic of an intrusion. Infected nodes may broadcast messages flooding the system with messages that are of no use and block the normal operation of the network.

In Figure 6 we observe that when in the upper graph there exist horizontal lines then we have a burst of traffic in the system. In the lower graph this burst is more easily observed. Traffic rate can be easily extracted from network trace files.

B. Attributes extracted

Attributes in the network traces datasets have all forms: continuous, discrete, and symbolic, with significantly varying resolution and ranges. Most pattern classification methods are not able to process data in such a format. Hence preprocessing was required before pattern classification models could be built. Pre-processing consistes of two steps: first step involved mapping symbolic-valued attributes to numeric-valued attributes and second step implemented scaling The

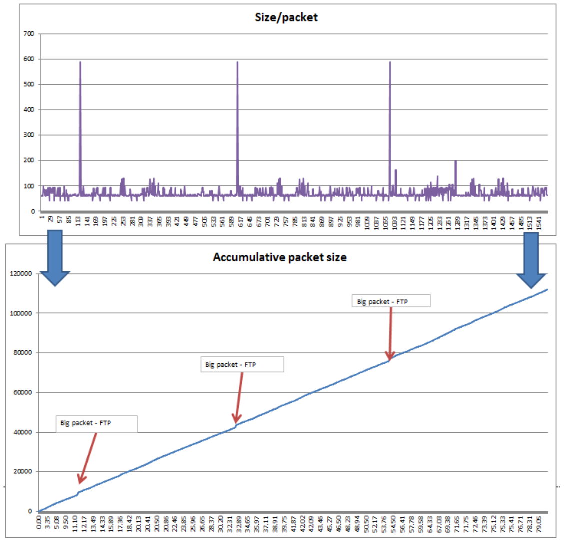


Fig. 5. Accumulative packet size

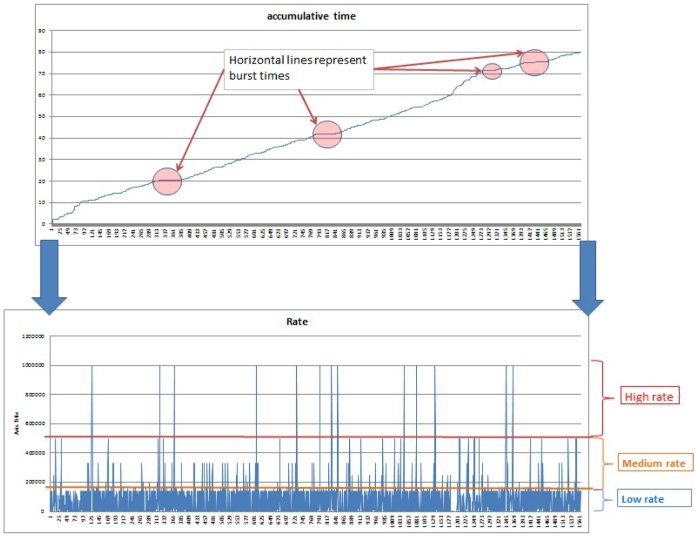


Fig. 6. Events over time - rate of traffic

main advantage of scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation.

Based on the above observations we used the network trace dataset in order to test our OCSVM module. The attributes that we used for this initial training / testing phase, were rate & packet size. The values were scaled to the range [0,1].

The rate (1^{st} attribute) was calculated using the equation 4:

$$Rate_{scaled} = \frac{Time\ difference}{Max\ time\ difference} \tag{4}$$

Time difference is calculated by the difference of time of current packet and the time of previous packet injected in the system.

The packet size (2^{nd} attribute) was scaled using the equation 5:

$$Packet_{scaled} = \frac{packet\ size}{Max\ packet\ size} \tag{5}$$

VI. Integration of OCSVM module

The above described OCSVM detection approach could be incorporated to the performance management scheme of the SCADA system. The performance management system contains sub systems and is summarized as follows:

- Firstly, the off line monitoring data is used to train the OCSVM and generate the model : Detector training
- Once tested, the model is transferred to performance monitoring and management system and detect the anomalies real time: Anomaly detection
- Once anomalies are detected they are classified into different classes according to their severities: Classification
- IDMEF files are send from the OCSVM module to the main correlator.

In order to cooperate with the other modules the OCSVM module needed to be integrated in the PID system and communicate with the other modules. Once an intrusion is detected several actions can be taken by the IDS (intrusion detection system. These actions include recording of intrusions in log files, sending of alert messages, limit the bandwidth of the intruder or even block all connections from the intruder. In order to better cooperate with the other components/modules that are being produced in the CockpitCI project the OCSVM model sends IDMEF [20] files.

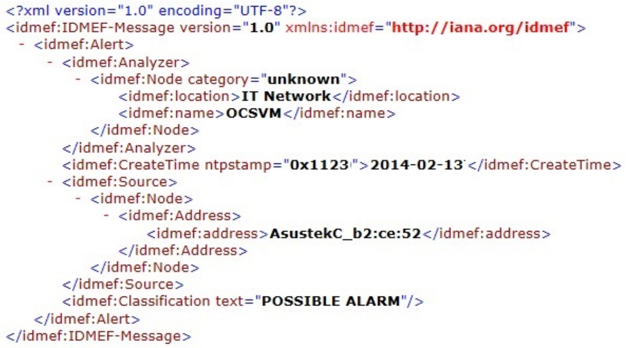
A. *Communication messages*

The IDMEF defines experimental standard for exchanging intrusion detection related events. As a standard, it can be used as a vendor or product independent enabling intercommunication between different agents such as NIDS or Honeypots. According to the RFC 4765 the data model of this format address several problems associated with representing intrusion detection alert data:

- As alert information is inherently heterogeneous, with some alerts with a very incomplete definition and others with very detailed information. The data model by using an object-oriented model provides extensibility via aggregation and sub classing.
- The intrusion detection environments are different. There are for example NIDS and HIDS analyzers. The data model defines support classes that accommodate the differences in the data sources among agents.
- The data model must allow for conversion to formats used by tools other than intrusion detection analyzers, for the purpose of further processing the alert information.
- The data model should accommodate the existing differences in operating environments and networks.

- Certain sensors deliver more or less information about certain types of attacks. The object oriented approach allows flexibility while the subclassing rule provides the integrity of the model.

A typical IDMEF file produced by our system is shown in Figure 7. The IDMEF message contains information about the source of the intrusion, the time of the intrusion detection, the module that detected the problem and a classification of the detected attack. The source node that the intrusion is detected is very important feature in an IDS system. Once the infected node is spotted the infection can be limited by the isolation of this node from the rest network. Fast and accurate detection of the source node of a contamination is crucial for the correct function of an IDS.



```
<?xml version="1.0" encoding="UTF-8"?>
<idmef:IDMEF-Message version="1.0" xmlns:idmef="http://iana.org/idmef">
 - <idmef:Alert>
    - <idmef:Analyzer>
       - <idmef:Node category="unknown">
            <idmef:location>IT Network</idmef:location>
            <idmef:name>OCSVM</idmef:name>
         </idmef:Node>
      </idmef:Analyzer>
      <idmef:CreateTime ntpstamp="0x1123">2014-02-13</idmef:CreateTime>
    - <idmef:Source>
       - <idmef:Node>
          - <idmef:Address>
               <idmef:address>AsustekC_b2:ce:52</idmef:address>
            </idmef:Address>
         </idmef:Node>
      </idmef:Source>
      <idmef:Classification text="POSSIBLE ALARM"/>
   </idmef:Alert>
</idmef:IDMEF-Message>
```

Fig. 7. Typical IDMEF message produced by OCSVM module

B. *OCSVM interfaces*

Once the detector training phase is complete, OCSVM detection is capable of detecting possible intrusions (abnormal behavior) on the SCADA system. Detection agents will gather new monitoring data (with corresponding attributes) and will feed the data to the OCSVM detection module. The detection module will classify each event whether it is a normal event or a possible intrusion. This information will then be send to the main correlator in order to react accordingly to the detected intrusions as shown in Figure 8

VII. PERFORMANCE EVALUATION

A. *Training of OCSVM model*

Training of OCSVM module was conducted using the transformed network trace file (Figure 9). To train the OCSVM, we adopt the RBF for the kernel equation. This kernel nonlinearly maps samples into a higher dimensional space so it can handle the case when the relation between class labels and attributes is nonlinear. The parameter σ is chosen 0.07 and the parameter ν 0.01.

The training model that is extracted after the training of the OCSVM is used for on line detection of malicious data. Since the model is based on features that are related to network traffic, and since the traffic of the system varies from area to area and from time period to time period, possible generation of multiple models could improve the performance of the module.

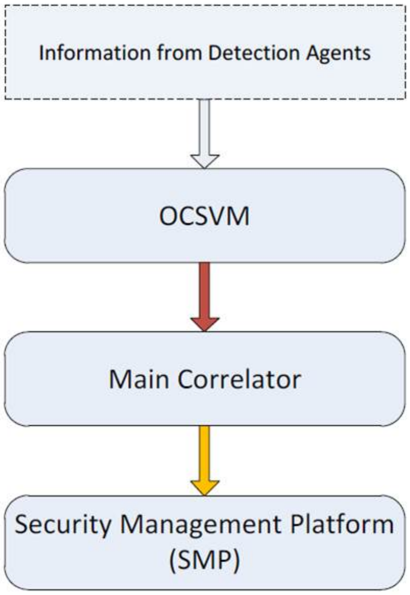
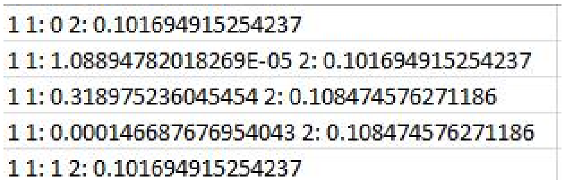

Information from Detection Agents
OCSVM
Main Correlator
Security Management Platform (SMP)

Fig. 8. Interfaces linked with OCSVM detection module



```
1 1: 0 2: 0.101694915254237
1 1: 1.08894782018269E-05 2: 0.101694915254237
1 1: 0.318975236045454 2: 0.108474576271186
1 1: 0.000146687676954043 2: 0.108474576271186
1 1: 1 2: 0.101694915254237
```

Fig. 9. Format of the transformed Network trace file

The network traffic in electric grids varies according to the activity which is not constant during the day. Also in some areas the activity follows different patterns according to the local demand. These characteristics maybe be critical for the proper training of the module and the accurate detection of intruders.

B. *Testing of OCSVM model*

In order to test our model we use the initial network trace file and we also spit the trace file in two separate files (A,B). The split was random and two datasets were constructed from the initial trace file. The two datasets were then used for training and testing. The dataset A was initially used for training and dataset B for testing and vice versa. The size of dataset A is 1000 rows and of B 570 rows. The results of our OCSVM detection module for each split are shown in Table VII-B. The accuracy of the classification of the data is high for all the tests conducted.

TABLE I. ACCURACY OF OCSVM MODULE UNDER DIFFERENT SPLIT OF DATA.

Split	*Accuracy*
All	98.8796
A	98.42
B	99.12

The outcomes of the classification method for the testing conducted in whole the dataset and in the two spitted sets are shown in Figure 11. The observed malicious data give small negative values and can only be classified as possible alerts. This is due to the fact that all the testing datasets are part of the initial trace file, which is captured during a normal operation of the system.

Malicious datasets that represent attack scenarios e.g. Man in the Middle (MITM) by ARP (Address Resolution Protocol) poisoning, SYN Flooding and honeypot [21] interaction, are used in the next subsection in order to test the performance of our $\mathcal{K}-\mathcal{OCSVM}$ module.

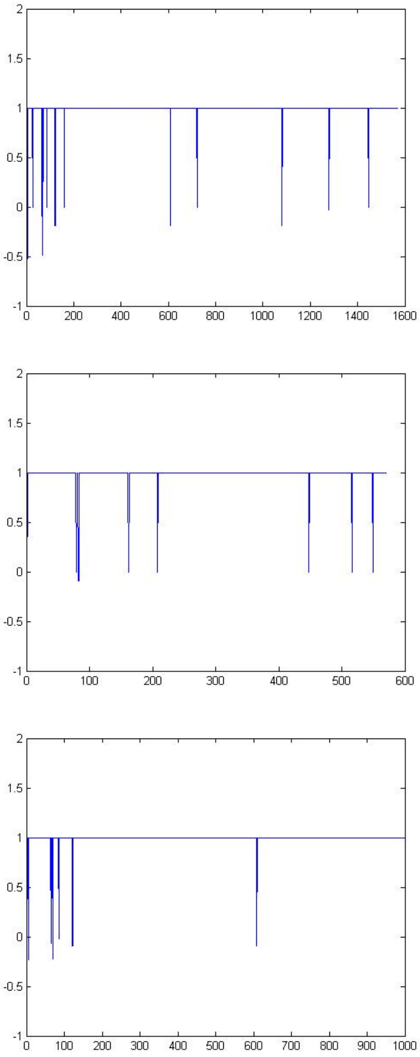


Fig. 10. OCSVM classification outcome for the three training / testing datasets (Upper figure:Original dataset, Middle figure: A/B dataset, Lower figure: B/A dataset

C. Testing of K-OCSVM model

We evaluated the performance of the method using data from the wireless network of the University campus and from a testbed that mimics a small-scale SCADA system. The parameters used for the evaluation of the performance of $\mathcal{K}-\mathcal{OCSVM}$ are listed in Table II.

Parameter	*Range of Values*	*Default value*
σ	0.1 - 0.0001	0.007
ν	0.002 - 0.05	0.01
$Threshold$	2-3	2

TABLE II. EVALUATION PARAMETERS

1) **Wireless network:** In order to test our model we use another network trace files sniffed from the wireless network. The testing trace file consists of 30.000 lines. We compare the performance of our proposed model against OCSVM classifiers having the same values for parameters σ and ν. We name each OCSVM classifier according to the parameters σ and ν : $OCSVM_{0.07,0.01}$ stands for OCSVM classifier with parameters $\sigma = 0.07$ and $\nu = 0.01$.

In Table III we show the number of observed anomalies detected from OCSVM and $\mathcal{K}-\mathcal{OCSVM}$ respectively. From this table it is shown how parameters σ, ν affect the performance of OCSVM. Even for a value of ν equal to 0.005, OCSVM produces almost 500 possible attacks, making the method inappropriate for a SCADA system where each false alarm is costly.

Parameter σ	*Parameter* ν	$\mathcal{K}-\mathcal{OCSVM}$	*ocsvm*
0.007	0.002	3	408
0.007	0.01	3	299
0.007	0.005	2	408
0.0001	0.01	3	274
0.1	0.01	2	295

TABLE III. PERFORMANCE EVALUATION OF $\mathcal{K}-\mathcal{OCSVM}$ AND OCSVM FOR $K_{thers} = 2$.

In figure 11 we present the outcome that OCSVM produces for the training network trace under different values of parameters σ and ν. From this figure it is obvious that the outcome is strongly affected by the values of these parameters, making $\mathcal{K}-\mathcal{OCSVM}$ necesary tool for proper intrusion detection.

2) **Testbed scenario:** The second trial is conducted off line with the use of two datasets extracted from the testbed (Figure 12. The testbed architecture mimics a small-scale SCADA system, comprising the operations and field networks and including a Human-Machine Interface Station (for process monitoring), a managed switch (with port monitoring capabilities, for network traffic capture), and two Programmable Logic Controller Units, for process control. The NIDS and OCSVM modules are co-located on the same host, being able to intercept all the traffic flowing on the network scopes.

During the testing period several attack scenarios are simulated in the testbed. These scenarios include network scan, network flood and MITM attack. Three kinds of attacks are being evaluated:

- **Network scan attack** In typical network scan attack, the attacker uses TCP/FIN scan to determine if ports are closed to the target machine. Closed ports answer with RST packets while open ports discard the FIN message. FIN packets blend with background noise on a link and are hard to be detected.
- **ARP cache spoofing - MITM attack ARP cache spoofing** is a technique where an attacker sends fake ARP messages. The aim is to associate the attacker's

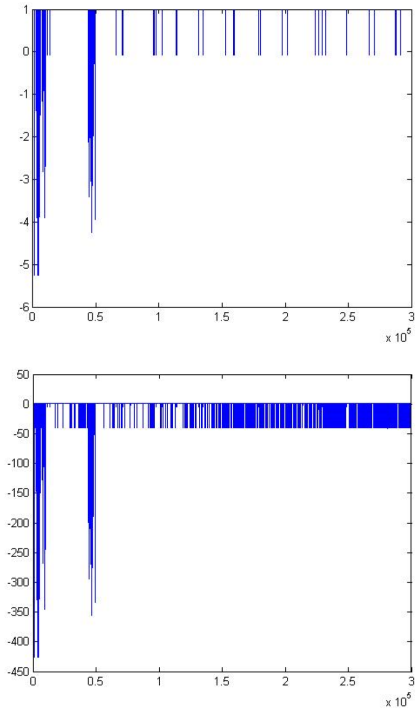


Fig. 11. OCSVM classification outcome for different values of parameters σ, ν Upper diagram : $OCSVM_{0.007,0.001}$, Lower diagram : $OCSVM_{0.01,0.05}$

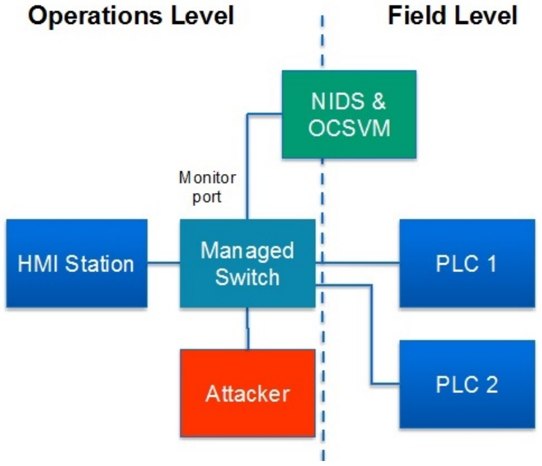


Fig. 12. Architecture of the testbed

MAC address with the IP address of another host, causing any traffic meant for that IP to be sent to attacker instead. The attacker could choose to inspect the packets, modify data before forwarding (**man-in-the-middle attack**) or launch a denial of service attack by causing some of the packets to be dropped.

- **DoS attack** Network flood is the instance where the attacker floods the connection with the PLC by sending SYN packets. In a TCP SYN flooding attack, an attackesends many SYN messages, with fictitious (spoofed) IP addresses, to a single node (victim). Although the node replies with SYN/ACK messages, these messages are never acknowledged by the client. As a result, many halfopen connections exist on the victim, consuming its resources. This continues until the victim has consumed all its resources, hence can no longer accept new TCP connection requests.

In Table IV we show the number of alert messages (IDMEF) sent from OCSVM and $\mathcal{K}-\mathcal{OCSVM}$ respectively. From this table it is shown how parameters σ, ν affect the performance of OCSVM for the testbed scenario. While for the same network trace file OCSVM produces from 10529 to 10704 alert messages according to the values of the parameters, $\mathcal{K}-\mathcal{OCSVM}$ produces the same 120 alert messages. All the reported attacks are concering the DoS attack that creates the biggest flunctuation in the network traffic.

Parameter σ	*Parameter* ν	$\mathcal{K}-\mathcal{OCSVM}$	*ocsvm*
0.007	0.002	120	10529
0.007	0.01	120	10703
0.007	0.005	120	10584
0.0001	0.01	120	10602
0.1	0.01	120	10704

TABLE IV. PERFORMANCE EVALUATION OF $\mathcal{K}-\mathcal{OCSVM}$ AND OCSVM FOR $K_{thers} = 2$.

3) **Testbed scenario with split testing periods:** Since the attacks are performed during different time periods we divide the testing dataset in several smaller ones, each containing a different attack. Testing data consists of normal data and attack data and the composition of the data sets are as follows:

- Testing set-A' : 1 - 5000: Normal data
- Testing set-B' : 5000 - 10000: Normal data + **Arp spoofing** attack + **Network scan**
- Testing set-C' : 10000 - 25000: Normal data + **Flooding Dos attack** + **Network scan**
- Testing set-D' : 25000 - 41000: Normal data + **MITM attack**

Dataset	*Initial alarms*	*Aggregated alarms*
A	129	2
B	658	3
C	9273	120
D	203	3
All	10507	3

TABLE V. AGGREGATED ALARMS PRODUCED BY $\mathcal{K}-\mathcal{OCSVM}$ ARE SIGNIFICANTLY DECREASED COMPARED TO THE INITIAL ALARMS

From table V we observe that not only the most important intrusions are detected and reported but also the total overhead on the system is limited. For all time periods the messages communicated reflect actual attacks in the network, except from the testing set-A'. In this time period HMI station demonstrated a significant variation in the rate that it injected packets in the system between testing and training of the module. This is due to the limited training of the OCSVM and can be avoided if training dataset consists of data that represent the traffic in the network during under work loads. The increased number of alarms created from $\mathcal{K}-\mathcal{OCSVM}$ for the dataset B' is due to the fact in this time period the attacker uses an exceshive number of SYN packets in order to flood the communication channel.

VIII. Conclusion

We have presented a intrusion detection module for SCADA systems that is based in OCSVM technique. The module is trained offline by network traces, after the attributes are extracted from the network dataset. The initial attributes used for training and testing of the module are rate and packet size which represent the traffic in the system. The intrusion detection module is part of an IDS system developed under CoCkpitCI. Output of the detection module is communicated to the system by IDMEF files that contain information about the source, time and severity of the intrusion.

After the execution of the $\mathcal{K}-\mathcal{OCSVM}$ method only severe alerts are communicated to the system by IDMEF files that contain information about the source, destination, protocol and time of the intrusion. The method is stable and its performance is not influenced by the selection of parameters ν and σ. The main feature of $\mathcal{K}-\mathcal{OCSVM}$ module is that it can perform anomaly detection in a time-efficient way, with good accuracy and low overhead. Low overhead is an important evaluation metric of a distributed detection module that is scattered in a real-time system, since frequent communication of IDMEF files from detection agents degrade the performance of the SCADA network. Recursive k-means clustering, reassures that small fluctuations on network traffic, which most of the times cause OCSVM to trigger false alarms, are ignored by the proposed detection module.

As future work we will conduct an in depth performance evaluation on proposed mechanism. Using malicious and attack-free datasets of the SCADA testbed, we are going to evaluate $\mathcal{K}-\mathcal{OCSVM}$'s performance in terms of false positive rate, accuracy and runtime. Using the evaluation outcomes we are planning to enhance the proposed $\mathcal{K}-\mathcal{OCSVM}$ in order to further decrease false alarms and improve overall performance. Additional attributes like reputation of the source and the protocol used for communication may add more precision to our system and it is a matter of future research. Use of different models according to the area or the time period may further improve the performance of the method.

Acknowledgment

The authors wish to acknowledge the financial support of the project CockpitCI, funded under European Framework-7 Programme (contract No. 285647).

References

[1] Y. Wang, J. Wong, and A. Miner, "Anomaly intrusion detection using one class svm," in *Information Assurance Workshop, 2004. Proceedings from the Fifth Annual IEEE SMC*. IEEE, 2004, pp. 358–364.

[2] K. Heller, K. Svore, A. D. Keromytis, and S. Stolfo, "One class support vector machines for detecting anomalous windows registry accesses," in *Workshop on Data Mining for Computer Security (DMSEC), Melbourne, FL, November 19, 2003*, 2003, pp. 2–9.

[3] J. Ma and S. Perkins, "Time-series novelty detection using one-class support vector machines," in *Neural Networks, 2003. Proceedings of the International Joint Conference on*, vol. 3, July 2003, pp. 1741–1745 vol.3.

[4] K.-L. Li, H.-K. Huang, S.-F. Tian, and W. Xu, "Improving one-class svm for anomaly detection," in *Machine Learning and Cybernetics, 2003 International Conference on*, vol. 5. IEEE, 2003, pp. 3077–3081.

[5] D. Dasgupta and F. A. Gonzalez, "An intelligent decision support system for intrusion detection and response," in *Information Assurance in Computer Networks*. Springer, 2001, pp. 1–14.

[6] A. Glazer, L. Michael, and S. Markovitch, "q-ocsvm: A q-quantile estimator for high-dimensional distributions," in *In Proceedings of The 27th Conference on Neural Information Processing Systems (NIPS-2013),Lake Tahoe, Nevada*, 2013.

[7] X. Song, G. Fan, and M. Rao, "Svm-based data editing for enhanced one-class classification of remotely sensed imagery," *Geoscience and Remote Sensing Letters, IEEE*, vol. 5, no. 2, pp. 189–193, 2008.

[8] L. Maglaras and J. Jiang, "Ocsvm model combined with k-means recursive clustering for intrusion detection in scada systems," in *Proceedings of the 10th Qshine conference*. EAI, 2014.

[9] E. Menahem, L. Rokach, and Y. Elovici, "Combining one-class classifiers via meta learning," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 2013, pp. 2435–2440.

[10] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[11] D. TAX, "One-class classification," *PhD thesis,Delft University of Technology, The Netherlands*, 2001.

[12] T. P. Runarsson and M. T. Jonsson, "Model selection in one-class ν-svms using rbf kernels," in *Proceedings of 16th International Congress and Exhibition on Condition Monitoring and Diagnostic Engineering Management*, 2003.

[13] X. Li, L. Wang, and E. Sung, "Adaboost with svm-based component classifiers," *Engineering Applications of Artificial Intelligence*, vol. 21, no. 5, pp. 785–795, 2008.

[14] R. R. R. Barbosa and A. Pras, "Intrusion detection in scada networks," in *Mechanisms for Autonomous Management of Networks and Services*. Springer, 2010, pp. 163–166.

[15] B. Zhu, A. Joseph, and S. Sastry, "A taxonomy of cyber attacks on scada systems," in *Internet of Things (iThings/CPSCom), 2011 International Conference on and 4th International Conference on Cyber, Physical and Social Computing*. IEEE, 2011, pp. 380–388.

[16] S. Karnouskos, "Stuxnet worm impact on industrial cyber-physical system security," in *IECON 2011-37th Annual Conference on IEEE Industrial Electronics Society*. IEEE, 2011, pp. 4490–4494.

[17] J. Jiang and L. Yasakethu, "Anomaly detection via one class svm for protection of scada systems," in *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2013 International Conference on*. IEEE, 2013, pp. 82–88.

[18] R. Zhang, S. Zhang, Y. Lan, and J. Jiang, "Network anomaly detection using one class support vector machine," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1, 2008.

[19] L. Maglaras and J. Jiang, "Intrusion detection in scada systems using machine learning techniques," in *Proceedings of the 2nd SAI conference*. SAI, 2014.

[20] H. Debar, D. A. Curry, and B. S. Feinstein, "The intrusion detection message exchange format (idmef)," 2007.

[21] L. Spitzner, "Honeypots: definitions and value of honeypots," *URL: http://www. t racking-hackers. com/papers/honeypots. html*, 2003.

Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming

Ahlam A. Sharief
Computer Science Department
Sudan University of Science and Technology
khartoum, Sudan
dreams585@yahoo.com

Alaa Sheta
Software Engineering Department
Zarqa University
Zarqa 13132, Jordan
asheta66@gmail.com

***Abstract*—Diabetes Mellitus is one of the deadly diseases growing at a rapid rate in the developing countries. Diabetes Mellitus is being one of the major contributors to the mortality rate. It is the sixth reason for death worldwide. Early detection of the disease is highly recommended. This paper attempts to enhance the detection of diabetic based on set of attributes collected from the patients to develop a mathematical model using Multigene Symbolic Regression Genetic Programming technique. Genetic Programming (GP) showed significant advantages on evolving nonlinear model which can be used for prediction. The developed GP model is evaluated using Pima Indian data set and showed higher capability and accuracy in detection and diagnosis of Diabetes.**

***Keywords*—Diabetes; Classification; Genetic Programming; Pima Indian data**

I. Introduction

Diabetes is one of the famous diseases that causing death. Based on measured statistics, it is the sixth reason for death worldwide. It was estimated that the world lose about 116 billion per year from medical care costs directly, and cost 580 billion indirectly (death, loss of work because of the deficit). Statistics showed that the high rates of deaths in developing countries are caused by diabetes disease. Early detection of the disease is highly recommended. It is essential to find a way that can help in early predicting this disease. A model with high accuracy, less complex and has efficient performance is urgently needed.

Diabetes Mellitus is simply caused by the failure of the body to produce the right amount of insulin to stabilize the amount of sugar in the body [1]. Most patients suffer this type of body failure are recommended to take insulin injection. This is called diabetes type I. Diabetes type II the patient body rejection to insulin. This type of patient is recommended to undergo certain health meal program as well as performing exercises to lose weight, plus taking oral medication. But heart diseases are likely to strike these patients in the long run [2].

Gestational Diabetes can occur temporarily during Pregnancy which is due to the hormonal changes and usually begins in the fifth or sixth month of pregnancy (between the 24th and 28th weeks). Gestational diabetes usually resolves once the baby is born. However, 25-50% of women with gestational diabetes will eventually develop diabetes later in their life, especially in those who require insulin during pregnancy and those who are overweight after their delivery.

A. General Diabetes Statistics

Due to the wide spread of type II infected diabetes in the USA, a survey was conducted by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) in collaboration with the American Diabetes Association [3], the result was 17.9 million have been diagnosed while 5.7 million are unaware that they are infected by the disease. Statistically 23.6 million people in America have been diagnosed type II diabetes positive. Table I show some statistics from the Gestational diabetes in the Middle East and Northern Africa [4]. Some reported statistics of infected people include:

- US women aged 20 years and older form 11.5 million which represent 10.2% of women in USA.
- US men aged 20 years and older form 12 million which represent 11.2% of men in USA.
- people under 20 years form 186,300.
- Adults over 60 years form 12.2 million.
- African Americans aged 20 years and older form 3.7 million (14.7% of all African Americans age 20 years and older).
- Hispanic/Latino Americans form 2.5 million (9.5% of all Hispanic/Latino Americans).
- Caucasian Americans aged 20 years and older form 14.9 million (9.8% of all Caucasian Americans age 20 years and older).

TABLE I. Gestational Diabetes Statistics

Country	**Extrapolated Incidence**	**Population Estimated**
Sudan	19.430	39, 148, 162
Iran	33.503	67.503.205
Iraq	12.594	25.374.691
Jordan	2.784	5.611.202
Kuwait	1.120	2.257.549
Lebanon	1.874	3.777.218
Saudi Arabia	12.803	25.795.938
Syria	8.942	18.016.874
UAE	1.252	2.523.915
Yemen	9.938	20.024.867
Egypt	37.778	76.117.421
Libya	2.795	5.631.585

The objective of this work is to explore the advantages of Multigene Symbolic Regression GP to classify the existence or non-existence of diabetic based on data collected from patient with various nature [5]. The proposed model can predict the class of the patient based on the eight attributes. The model is based on number of measured features of the patients. They include: the number of times pregnant, the results of an oral glucose tolerance test, diastolic blood pressure (mm/Hg), E-Triceps skin fold thickness (mm), 2-h serum insulin (micro U/ml), body mass index, diabetes pedigree function, Age (year).

The paper is structured as follows. In section II, we provide a literature review on the basic research work in the area of diabetic research based on soft computing techniques such as Artificial Neural Networks. In section III, basic process of GP is described. The expansion of Multigene Symbolic Regression GP approach is provided in section IV. The developed results are presented in section VI including the inputs and output of the model, the experimental setup and the developed mathematical GP Model. Finally we introduce the conclusion and future work.

II. Literature Review

The need for an accurate predictor for the diabetes is highly needed. Not only this, but also a predictor that is extremely automated and with less human interference. A diabetic predictor should meet the following specification; efficient modeling, applicability and accuracy and be trusted. It should be compatible with various diagnostic techniques.

Many prediction techniques are used, but the Multi-layer Percepton (MLP) is the most common [6]–[8]. ANN consists of fully connected layers. In the training phase of the prediction, the learning algorithm examines the inputs. While during the testing phase, it examines the outputs and the other unexamined parts during the training phase.

Anthropometrical Body surface scanning data was used to construct a classification model for diabetes type II in [9]. The model applies four data mining approaches. This model is meant to select and point out the appropriate and necessary decision tree for classifying diabetic diseases. It incorporates Artificial Neural Network, Decision Tree, Logistic Regression and Rough sets. In [10] authors used the classification tree for the classification and regression with a binary target. It introduces ten attributes including age, sex, emergency department visits, comorbidity index, dyslipidemia, hypertension, cardiovascular disease, and retinopathy and end-stage renal disease. The cascade learning system which is based on generalized discriminant analysis was introduced by [11]. It has also linked the system with the least square support vector machine in order to perform the classification of diabetes diseases. This uses the classification accuracy, k-fold cross-validation method and confusion matrix.

A method to discover key attributes affecting diabetic diseases was introduced in [12]. The method is called feature selection method. Then it introduced the three classification complementary techniques including Naive Bayes and C4.5. In [13] authors developed and upgraded the Linear Discriminant Analysis (LDA) and integrated it into the automatic diagnosis system. All these models functions primarily in the area of classification. But this method is meant to be accurate and well performed.

The fuzzy approaches have recently become the well-known approaches for improving classification models. Fuzzy Neural Networks (FNNs) and artificial neural networks have been recently integrated hybrid classification model that helps well in diagnosing and classifying the state of the diabetic diseases. This model was presented by [14]. Multi-objective genetic programming approach is proposed by [15] to develop Pareto optimal decision trees in diabetes classification. In [16], GP was used to generate new features by making combinations of the existing diabetes features.

III. Genetic Programming

GP works on a population of individuals, each of which represents a potential solution to a problem. GP was introduced by J. Koza in 1992 at Stanford. A flow chart for GP evolutionary process is shown in Figure 1. In order to solve a problem, it is necessary to specify the following [17]:

- **The terminal set:** A set of input variables or constants.
- **The function set:** A set of domain specific functions used in conjunction with the terminal set to construct potential solutions to a given problem. For symbolic regression this could consist of a set of basic mathematical functions, while Boolean and conditional operators could be included for classification problems.
- **The fitness function:** Fitness is a numeric value assigned to each member of a population to provide a measure of the appropriateness of a solution to the problem in question.
- **The termination criterion:** This is generally a predefined number of generations or an error tolerance on the fitness.

In order to further illustrate the coding procedure and the genetic operators used for GP, a symbolic regression example will be used. Consider the problem of predicting the numeric value of an output variable, y, from two input variables a and b. One possible symbolic representation for y in terms of a and b would be: $y = \frac{a-b}{3}$.

Figure 2 demonstrates how this expression may be represented as a tree structure. With this tree representation, the genetic operators of crossover and mutation must be posed in a fashion that allows the syntax of resulting expressions to be preserved. Figure 3 shows a valid crossover operation where the two parent expressions are given in Equations 1 and 2. The two offspring are given in Equation 3 and 4. Parent 1 (y^1) and Parent 2 (y^2) are presented in Equations 1 and 2. The developed offspring 1 (y^3) and offspring 2 (y^4) are presented in Equation 3 and 4.

$$y^1 = \frac{a-b}{3} \tag{1}$$

$$y^2 = (c-b) \times (a+c) \tag{2}$$

$$y^3 = \frac{a-b}{a+c} \tag{3}$$

$$y^4 = (c-b) \times 3 \tag{4}$$

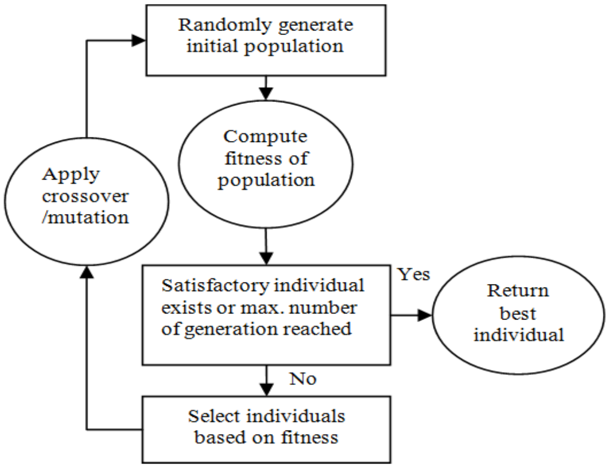

Randomly generate initial population
Compute fitness of population
Apply crossover /mutation
Satisfactory individual exists or max. number of generation reached
Yes
Return best individual
No
Select individuals based on fitness

Fig. 1. Flow chart of the GP algorithm [18]

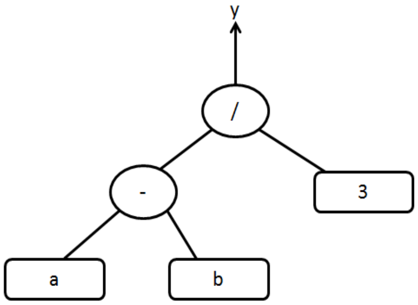


Fig. 2. Representation of a numeric expression using tree structure

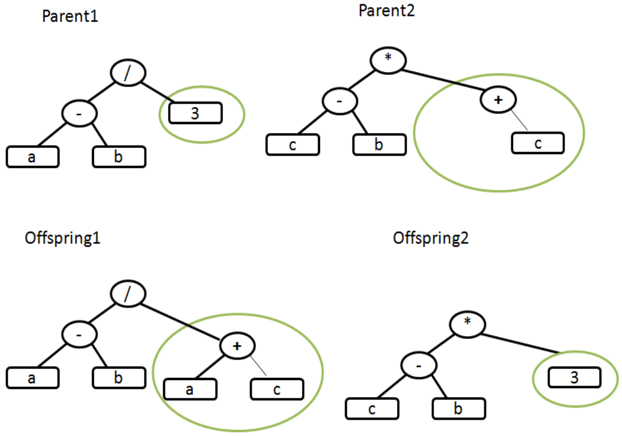


Fig. 3. A typical crossover operation

IV. Multigene Symbolic Regression GP

Typically, symbolic regression is performed by using GP to evolve a population of trees, each of which encodes a mathematical equation that predicts $n \times 1$ vector of outputs y using a corresponding $n \times m$ matrix of inputs X where N is the number of observations of the response variable and M is the number of input (predictor) variables [17]. In contrast, in Multigene symbolic regression each symbolic model (and each member of the GP population) is a weighted linear combination of the outputs from a number of GP trees, where each tree may be considered to be a gene [19]. For example, the Multigene model shown in Figure 4 predicts an output variable using input variables x_1, x_2, x_3. This model structure contains non-linear terms (e.g. the hyperbolic tangent) but is linear in the parameters with respect to the coefficients $\alpha_0, \alpha_1, \alpha_2$.

In practice, the user specifies the maximum number of genes G_{max} and the maximum tree depth D_{max} therefore an exert can control the model complexity. In particular, we have found that enforcing stringent tree depth restrictions (i.e. maximum depths of 4 or 5 nodes) often allows the evolution of relatively compact models that are linear combinations of each model, the linear coefficients are estimated from the training data using ordinary least squares techniques.

Hence, Multigene GP combines the power of classical linear regression with the ability to capture non-linear behavior without the need to pre-specify the structure of the non-linear model. In [20] it was shown that Multigene symbolic regression can be more accurate and computationally efficient than the standard GP approach for symbolic regression.

Here, the first parent individual contains the genes (G_1 G_2 G_3) and the second contains the genes (G_4 G_5 G_6 G_7) where G_{max} equals to 5. Two randomly selected crossover points are created for each individual. The genes enclosed by the crossover points are denoted by [ ].

$$(G_1 \ [\ G_2\ ]\ G_3)\ (G_4\ [\ G_5\ G_6\ G_7\ ])$$

The genes enclosed by the crossover points are then exchanged resulting in two new individuals as follows:

$$(G_1\ G_5\ G_6\ G_7\ G_3)\ (G_4\ G_2)$$

Two point high level crossover allows the acquisition of new genes for both individuals but also allows genes to be removed. If an exchange of genes results in an individual containing more genes than G_{max} then genes are randomly selected and deleted until the individual contains G_{max} genes.

The user can set the relative probabilities of each of these recombination processes. These processes are grouped into categories called events. The user can then specify the probability of crossover events, direct reproduction events and mutation events. These must sum to one. The user can also specify the probabilities of event subtypes, e.g. the probability of a two point high level crossover taking place once a crossover event has been selected, or the probability of a sub tree mutation once a mutation event has been selected.

An example of Multigene model is shown in Figure 4. The presented model can be introduced mathematically as given in Equation 5. GPTIPS Matlab Toolbox provides default values for each of these probabilities so the user does not need to explicitly set them [21].

$$\alpha_{\mathbf{0}} + \alpha_{\mathbf{1}}(0.41x_1 + tanh(x_2x_3)) + \alpha_{\mathbf{2}}(0.45x_3 + \sqrt{x_2}) \quad (5)$$

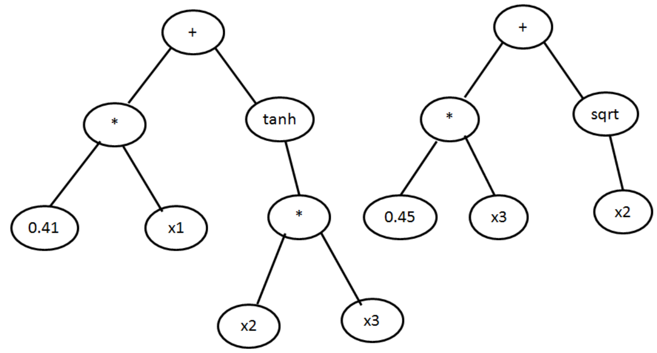


Fig. 4. Example of a Multigene Symbolic Regression Model

V. Performance Criterion

Number of evaluation criterion was computed to evaluate the performance of the developed models. The Route Mean Square (RMS) was used as the fitness function for genetic programming. RMS can be described by Equation 6.

$$RMS = \sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2} \quad (6)$$

Other performance criterion was used to evaluate the goodness of the developed GP model. The set of criterion used are given as follows:

- Sensitivity (Sens):

$$Sens = \frac{TP}{TP + FN} \quad (7)$$

- Specificity (Spec):

$$Spec = \frac{TN}{FP + TN} \quad (8)$$

- Positive Predicted Value (PPV):

$$PPV = \frac{TP}{TP + FP} \quad (9)$$

- Negative Predicted Value (NPV):

$$NPV = \frac{TN}{FN + TN} \quad (10)$$

- Accuracy (Acc):

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

Given that:

- True Positive (TP): Sick people correctly diagnosed as sick.
- False Positive (FP): Healthy people incorrectly identified as sick.
- True Negative (TN): Healthy people correctly identified as healthy.
- False Negative (FN): Sick people incorrectly identified as healthy.

VI. Developed Results

A. Model Inputs and Output

Pima Indian is a homogeneous group that inhabits the area around American, but they are popular for being the most infected group with type II diabetes. Pima Indians diabetes data can even be retrieved from UCI Machine Learning Repository's web site [5]. So, they are subject of intense studies in type II diabetes. The data consist of eight input variables and one output (0,1). The GP mathematical model has the inputs and output presented in Table II. We used 500 samples as a training set and 100 samples as a testing set. The data set was normalized according to Equation 12.

$$x^{new} = \frac{x^{old} - x_{min}}{x_{max} - x_{min}} \quad (12)$$

x_{max} and x_{min} are the maximum and minimum values of the array x, respectively. x^{new} is the newly computed value based on the value of x^{old}.

TABLE II. Inputs and Output for Diabetic Prediction Model

Inputs	The number of times pregnant	x_1
	The results of an oral glucose tolerance test	x_2
	Diastolic blood pressure (mm/Hg)	x_3
	E-Triceps skin fold thickness (mm)	x_4
	2-h serum insulin (micro U/ml)	x_5
	Body mass index	x_6
	Diabetes pedigree function	x_7
	Age (year)	x_8
Output	Predicted class	y

B. Experimental Setup

In this research, we adopted a GPTIPS toolbox [21] to develop our results. In GPTIPS, the initial population is constructed by creating individuals that contain randomly generated GP trees with between 1 and G_{max} genes. During the run, genes are acquired and deleted using a tree crossover operator called two point high level crossover. This allows the exchange of genes between individuals and it is used in addition to the standard GP recombination operators.

Some parameters have to be defined by the user at the beginning of the evolutionary process. They include: population size, probability of crossover, mutation probability and the type of the selection mechanism. User has also to setup the maximum number of genes G_{max} where a model is allowed to have. The maximum tree depth D_{max} allows us to change the complexity of the evolved models. Restricting the tree depth helps evolving simple model but it may also reduce the performance of the evolved model.

A prior knowledge on the problem domain helps in designing a function set which could speed up the evolutionary process for model development. The adopted function set to develop the GP model is given as:

$$F = \{+, -, \times\}$$

TABLE III. GP TUNING PARAMETERS

Population size	100
Number of generation	100
Selection mechanism	Tournament
Max. tree depth	7
Probability of Crossover	0.85
Probability of Mutation	0.1
Max. No. of genes allowed in an individual	6

Crossover was performed with the two-point high-level crossover operator. Once the two parent individuals have been selected, two gene crossover points are selected within each parent. Then the genes enclosed by the crossover points are swapped between parents to form two new offspring.

C. Developed Mathematical GP Model

The data set described earlier was loaded then the Multigene GP was applied using GPTIPS Tool. The parameters of the algorithm were tuned as listed in Table III. In Figure 5, we show the convergence of GP over 100 generations. It can be clearly seen that the final model is a simple and compact mathematical model which is easy to evaluate. The performance measurements for the model was computed and summarized in Table V. The best generated diabetic prediction Multigene GP model is given in Table IV.

TABLE V. PERFORMANCE OF THE GP MODEL WITH $x_1, \ldots, x_8$ AS INPUTS

Criteria	Training	Testing
Sensitivity	0.90881	0.95946
Specificity	0.56593	0.69231
Accuracy	0.78533	0.89873
Positive Predicted Value	0.7803	0.85714
Negative Predicted Value	0.784	0.890

We also explored the idea of considering a subset of the features used to develop the GP model. Thus, we considered the features x_3, x_6 and x_8 to develop the output class y of diabetic type. Running GP with a population size 30 and 100 generations with the same tuning parameters such as tree depth, maximum number of genes, probability of crossover and probability of mutation we produced the results in this case. The performance measurements for the developed GP model was computed and summarized in Table VI. In Figure 6, we show the convergence of GP in the case with less number of features. The developed GP model is presented in Table VII.

TABLE VI. PERFORMANCE OF THE GP MODEL WITH x_3, x_6 AND x_8 AS INPUTS

Criteria	Training	Testing
Sensitivity	0.84591	0.81081
Specificity	0.47253	0.65385
Accuracy	0.73699	0.86957
Positive Predicted Value	0.63704	0.54839
Negative Predicted Value	0.71	0.77

VII. CONCLUSIONS AND FUTURE WORK

In this paper, a GP mathematical model was developed to provide a solution to the diabetic problem. The developed model was able to classify patient type. The developed classification accuracy obtained based on Multigene GP is high with respect to sensitivity, specificity, accuracy, positive predicted and negative predicted values. These evaluation criterions proved that Multigene GP is beneficial for diabetic patient classification. The knowledge gained is comprehensible and can enhance the decision making process by the physician. We plan to expand this research to detect the most significant attributes which indicate diabetic.

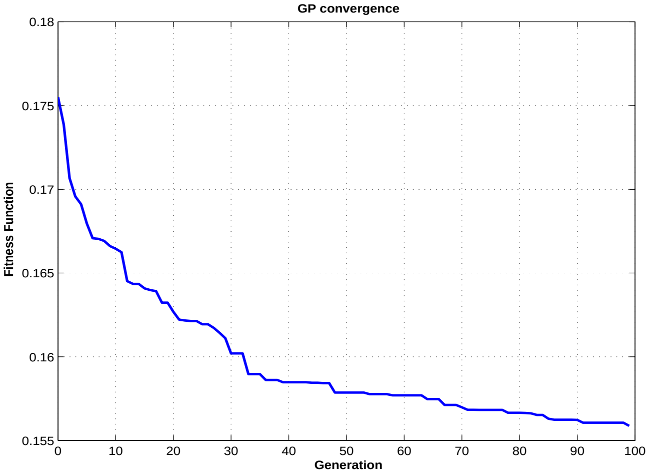


Fig. 5. Convergence of the GP evolutionary process

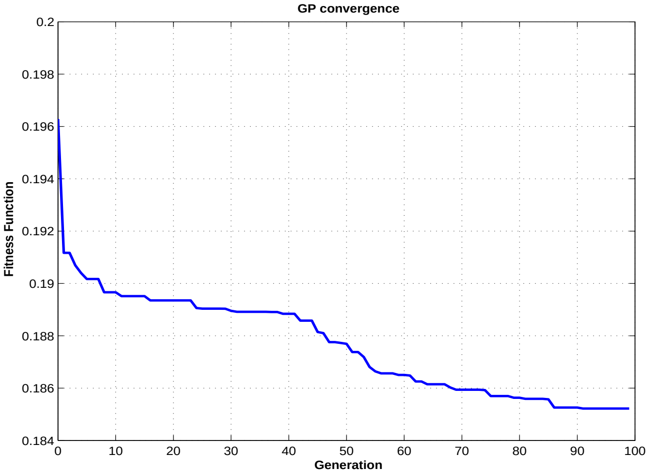


Fig. 6. Convergence of the GP evolutionary process

REFERENCES

[1] B. M. Patil, R. C. Joshi, and D. Toshniwal, “Hybrid prediction model for type-2 diabetic patients,” *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8102–8108, 2010.

[2] C. Zecchin, A. Facchinetti, G. Sparacino, G. De Nicolao, and C. Cobelli, “A new neural network approach for short-term glucose prediction using continuous glucose monitoring time-series and meal information,” *Proceedings of the IEEE International Conference on Engineering Med. Biol. Soc.*, pp. 5653–6, 2011.

[3] G. E. C. Estrada, L. del Re, and E. Renard, “Nonlinear gain in online prediction of blood glucose profile in type 1 diabetic patients,” in *CDC*, pp. 1668–1673, 2010.

[4] USA, “US census bureau, population estimates, statistics by country for gestational diabetes,” 2004.

TABLE IV. A GP MODEL WITH INPUTS: $x_1, \ldots, x_8$

$$\begin{aligned} y &= 0.3636 * x_1 - 0.658 * x_2 - 0.4626 * x_4 - 0.4626 * x_5 - 0.3636 * x_3 * x_6 + 1.349 * x_2^2 + 1.331 * x_6^2 \\ &+ 0.4626 * x_2 * (x_4 + 2 * x_5) * (-x_6^2 + x_3 + x_4) + 0.658 * x_2 * x_5 * x_7 \\ &- 0.3636 * x_2 * (x_3 + x_4) * (x_1 - x_5) - x_2^2 * x_7 * (3.031 * x_2 - 3.031) \\ &+ 12.38 * x_7 * (x_4 + x_5) * (x_1 - x_5) * (x_3 - x_4) - 0.06898 \end{aligned}$$

TABLE VII. A GP MODEL WITH INPUTS: x_3, x_6 AND x_8

$$\begin{aligned} y &= 0.3748 * x_6^2 * (x_3 * (x_6 + x_8) * (x_3^2 + x_8) + 2.66) * (x_8 - x_6 + 2.559) - 1.435 * (x_3 - x_8) * (x_6 - x_8) \\ &- 0.03142 * x_6 * ((x_6 - x_8) * (2 * x_6 + x_8) + 2 * x_3 * x_6 * x_8 * (x_6 + 5.125)) * (x_8 - x_6 * (x_3 - x_6)) * (2.559 * x_6 - 2.559 * x_8 + 69.97 * x_6 * x_8) \\ &+ 0.02144 \end{aligned}$$

[5] M. W. Aslam and A. K. Nandi, “Detection of diabetics using genetic programming,” in *European Signal Processing Conference*, no. 18, (Aalborg, Denmark), August 2010.

[6] L. M. Silva, J. M. de Sá, and L. A. Alexandre, “Data classification with multilayer perceptrons using a generalized error function,” *Neural Networks*, vol. 21, no. 9, pp. 1302–1310, 2008.

[7] A. Elkamel, S. Abdul-Wahab, W. Bouhamra, and E. Alper, “Measurement and prediction of ozone levels around a heavily industrialized area: a neural network approach,” *Advances in Environmental Research*, vol. 5, no. 1, pp. 47 – 59, 2001.

[8] R. S. Selvaraj, K. Elampari, R. Gayathri, and S. J. Jeyakumar, “A neural network model for short term prediction of surface ozone at tropical city,” *International Journal of Engineering Science and Technology*, vol. 2, no. 10, pp. 5306–5312, 2010.

[9] C.-T. Su, C.-H. Yang, K.-H. Hsu, and W.-K. Chiou, “Data mining for the diagnosis of type ii diabetes from three-dimensional body surface anthropometrical scanning data,” *Computers & Mathematics with Applications*, vol. 51, no. 6-7, pp. 1075–1092, 2006.

[10] J. T. Tennis, “Three spheres of classification research: Emergence, encyclopedism, and ecology,” in *ASIS SIG/CR Classification Research Workshop*, 2002.

[11] K. Polat, S. Güneş, and A. Arslan, “A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine,” *Expert Syst. Appl.*, vol. 34, pp. 482–487, Jan. 2008.

[12] Y. Huang, P. J. McCullagh, N. D. Black, and R. Harper, “Feature selection and classification model construction on type 2 diabetic patients’ data,” *Artificial Intelligence in Medicine*, vol. 41, no. 3, pp. 251–262, 2007.

[13] D. Çalisir and E. Dogantekin, “An automatic diabetes diagnosis system based on lda-wavelet support vector machine classifier,” *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8311–8315, 2011.

[14] H. Kahramanli and N. Allahverdi, “Design of a hybrid system for the diabetes and heart diseases,” *Expert Systems with Applications*, vol. 35, no. 12, pp. 82 – 89, 2008.

[15] E. Mugambi and A. Hunter, “Multi-objective genetic programming optimization of decision trees for classifying medical data,” in *Knowledge-Based Intelligent Information and Engineering Systems* (V. Palade, R. Howlett, and L. Jain, eds.), vol. 2773 of *Lecture Notes in Computer Science*, pp. 293–299, Springer Berlin Heidelberg, 2003.

[16] M. W. Aslam, Z. Zhu, and A. K. Nandi, “Feature generation using genetic programming with comparative partner selection for diabetes classification,” *Expert Systems with Applications*, vol. 40, no. 13, pp. 5402 – 5412, 2013.

[17] J. R. Koza, *Genetic programming: on the programming of computers by means of natural selection.* Cambridge, MA, USA: MIT Press, 1992.

[18] A. F. Sheta, H. Faris, and E. Öznergiz, “Improving production quality of a hot-rolling industrial process via genetic programming model,” *Int. J. Comput. Appl. Technol.*, vol. 49, pp. 239–250, June 2014.

[19] J. Koza, “Evolving a computer program to generate random numbers using the genetic programming paradigm,” in *Proceedings of the Fourth International Conference on Genetic Algorithms*, Morgan Kaufmann, La Jolla,CA, 1991.

[20] M. Hinchliffe, H. Hiden, B. McKay, M. Willis, M. Tham, and G. Barton, “Modelling chemical process systems using a multi-gene genetic programming algorithm,” in *Late Breaking Papers at the Genetic Programming 1996 Conference Stanford University July 28-31, 1996* (J. R. Koza, ed.), (Stanford University, CA, USA), pp. 56–65, Stanford Bookstore, 28–31 July 1996.

[21] D. P. Searson, D. E. Leahy, and M. J. Willis, “GPTIPS : An open source genetic programming toolbox for multigene symbolic regression,” in *Proceedings of the International Multi-conference of Engineers and Computer Scientists 2010 (IMECS 2010)*, vol. 1, (Hong Kong), pp. 77–80, 17-19 Mar. 2010.