A Proposed Hybrid Effective Technique for Enhancing Classification Accuracy

Ibrahim M. El-Hasnony
Information Systems Department,
Faculty of Computer & Information Sciences,
Mansoura University,
Mansoura, EGYPT

Omar H. Al-Tarawneh
Information Technology Department,
Al-Zahra College for Women,
Muscat,
Oman

Hazem M. El-Bakry
Information Systems Department,
Faculty of Computer & Information Sciences,
Mansoura University,
Mansoura, EGYPT

Mona Gamal
Information Systems Department,
Faculty of Computer & Information Sciences,
Kafer el sheikh University,
Kafer el sheikh,
EGYPT

***Abstract*—The automatic prediction and detection of breast cancer disease is an imperative, challenging problem in medical applications. In this paper, a proposed model to improve the accuracy of classification algorithms is presented. A new approach for designing effective pre-processing stage is introduced. Such approach integrates K-means clustering algorithm with fuzzy rough feature selection or correlation feature selection for data reduction. The attributes of the reduced clustered data are merged to form a new data set to be classified. Simulation results prove the enhancement of classification by using the proposed approach. Moreover, a new hybrid model for classification composed of K-means clustering algorithm, fuzzy rough feature selection and discernibility nearest neighbour is achieved. Compared to previous studies on the same data, it is proved that the presented model outperforms other classification models. The proposed model is tested on breast cancer dataset from UCI machine learning repository.**

***Keywords*—Data mining; bioinformatics; fuzzy rough feature selection; correlation feature selection and data classification**

I. Introduction

Medical data characterized by being intricate, noisy and immense so there are challenges in decision making for patient health. Therapeutic datasets contains details related to patients, past diagnosis, treatment cost etc. therefore new approaches to extract and analyse valuable information from such data are required. These approaches improve decision making in regards to patient treatment. Because of the significant increase in digital data, good exploration and analysis of data is needed. Healthcare data that stored digitally is about 500 petabytes worldwide in 2012 and in 2020 expected to reach 25000 petabytes [6]. Breast cancer is one of the largest reasons for cancer deaths among women. Early expectation of the trademark of bosom protuberances (benign or malignant) happening in patients accordingly help to focus a suitable treatment for the cancer. Extracting valuable information from the breast cancer therapeutic datasets may help in early expectation of the disease.

Data mining concept is the methodology of extracting knowledge or finding models from huge amount of data. Data mining can be called ‘Knowledge mining from data [1]. Predictive data mining will be utilized to forecast some property of incoming data, for example how to classify it. The Data Mining (DM) techniques provide efficient methods to improve statistical tools for future pattern forecasting [2]. The idea for the developments of this advanced analysis is to extract useful information from large datasets and transforming it to meaningful pattern or structure that can be utilized later. The methods involved in that step are machine learning, database systems, artificial intelligence, statistics and business intelligence. Data mining is about making solutions by analysing data that are presented in datasets [3]. Because of academic and industrial models development there are need for hybrid models development such models represented (Fig. 1) by Cios et al. [4].

Data mining provides technologies that can be used in many organizations. Health awareness organizations can utilize these advances for characterizing patients that have comparative highlights and propose successful treatment.

Mining get to be imperative in healthcare management in light of the fact that they require techniques for effective examination to distinguish important, hidden and valuable data from medical data sets. Data classification issues in healthcare services result from the instability and high dimensionality nature of gathered medical information. In therapeutic frameworks, data mining can be utilized to attain fascinating advantages like lower expense answer for patient, discovering fraud in health insurance, looking for reasons for different diseases and discovering medical treatment solutions for them [5].

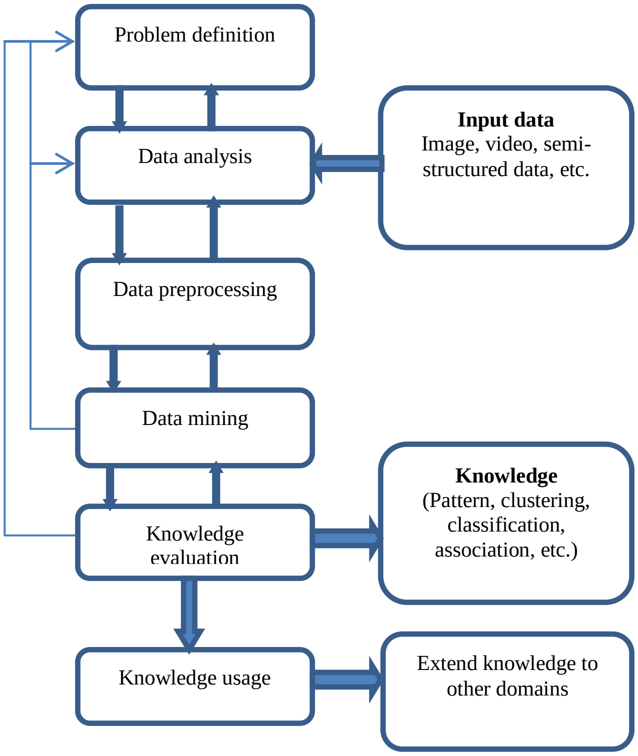
Problem definition
Data analysis
Input data
Image, video, semi-structured data, etc.
Data preprocessing
Data mining
Knowledge
(Pattern, clustering, classification, association, etc.)
Knowledge evaluation
Knowledge usage
Extend knowledge to other domains

Fig. 1. Knowledge discovery process.

This paper proposes a hybrid model that represents a unified schema for the classification algorithms. The model studies how the efficient preparation and selection of data participate in the improvement of classification algorithms. The model is a hybrid of the pre-processing phase and the classification phase. The pre-processing phase intends to cluster the data set using the K-mean clustering algorithm then use fuzzy rough feature selection (FRFS) or correlation feature selection (CFS) on each cluster to produce the reduct. The reducts of the clusters are then combined to produce the final set of features used for classification later. The discernibility nearest neighbour algorithm is trained by the reduced data set and classifies unseen cases of the test data. The accuracy of the hybrid model is judged by the 10 folds cross validation and percentage split by 80-20 for training and test respectively. The experimental results were applied using the WEKA and Rapid Miner data mining tools. The results showed that the proposed model improved the classification accuracies by 2.92% more than the old models that may classify original data or be applied on reduced data.

The rest of the paper is organized as follows: Section II is a quick review on the previous studies in using data mining techniques for medical applications and classification of breast cancer data set. Section III describes the techniques and methodologies used in this study such as data classification, data clustering and feature reduction algorithms. Section IV presents an overview on the proposed system and its modules. Experimental results and conclusion will display in Sections V and VI, respectively.

II. Related Work

Because biomedical is considered important and critical issue, many research papers seek to enhance medical data classifications accuracy. Agarwal and Pandey [7] performed a comparative study between different machine learning techniques such as fuzzy inference systems (FIS), perceptron neural networks and backprobagation neural networks. Their experiment tries to eye clinic datasets by using matlab simulation and proved that perceptron neural networks are better in its results from other techniques. From their study, also perceptron neural networks is simple and fuzzy logic and back propagation neural networks are widely used in several research area, they didn’t produce good results for the data

Anushya and Pethalakshmi [8] used fuzzy logic because of their comprehensible result for evaluating the accuracy of occurrence of a heart disease with several data mining classification techniques such as decision tree, k-means, naïve Bayes and neural networks. Authors used the classifiers to classify heart dataset as healthy or sick. They used sensitivity with specificity to measure the accuracy of classifiers. The results showed better accuracy in using fuzzy logic with k-means classifier but with the whole dataset without reducing its features.

Rawat and Burse [9] proposed a soft computing genetic-neuro fuzzy system for medical data mining diagnosis. They used genetic algorithms for feature selection combined with adaptive neuro-fuzzy inference system (ANFIS) for classification. Data grouped from UCI to ovarian cancer data. The system achieved higher accuracy with minimum cost. Cost decreased when using genetics for reducing data.

Shukla and Agarwal [10] presented hybrid system of combining clustering with classification with some k-means improvement. The system was tested against Tuberculosis Dataset .This model starts by handling data pre-processing and feature selection using principal components analysis (PCA). Then it applies clustering using modified k-means and comparing classification accuracy with three classifiers (Naïve Bayes, Decision Tree and Artificial Neural Networks (ANN)). The model proved that the modified K-means results are better than using the original K-means. Hamdan and Garibaldi [11] proposed a framework for survival modelling by using ANFIS fuzzy inference system. This framework consists of pre-processing data against missing values by replacing or discarding it with respect to data types and volumes then used ANFIS for implementation and using dataset related to operative surgery for ovarian cancer patients. They proved the predictive power of proposed framework and facilitation for clinician to understand the process of data by set of linguistic rules. Cedeño et al. [12] presented a novel enhancement in neural network training for pattern classification. The proposed training algorithm is roused by the biological met plasticity property of neurons and Shannon’s information theory. Joshi et al. [13], the outcome of their research is justified that clustering by k-means algorithm and FF algorithm are useful for early diagnosis of the breast cancer patients.

Many of previous studies interested in using powerful and intelligent classification algorithms for their work. The proposed model concentrates on pre-processing and not just for data reduction but selecting the features that have role in improving classification accuracies that help in prediction model specially the field about healthcare.

III. Materials and Methods

A. Methodology

In this paper the proposed system focus first on applying pre-processing on medical data such as handling missing data, clustering data, data reduction, performing comparison study between different algorithms and measuring the contribution of each method to improve the performance. Second, different techniques that can be used for optimizing classification accuracy of bioinformatics data have applied. The proposed system applies different data mining and artificial intelligence techniques.

B. Data Pre-Processing

Medical data is not complete and need several pre-processing steps that are performed by several techniques [14], [15]. Machine learning fields have several model analysis, design and data pre-processing techniques that guaranty high performance in achieving accuracy in its results. There are many problems in biomedical and medicine research that can make use of machine learning techniques in its tasks [16].

The pre-processing stages are very important and critical issue to ensure the success of data mining and data warehouse in time and space. Medical data is incomplete, noisy, and inconsistent. There are many different ways to solve such problems. Data pre-processing include several methods such as data reduction and data cleaning [4], [17].

1) Data Reduction: A central issue in machine learning is recognizing specific set of features from which a classification model can be built. Data reduction can be used to reduce the data set achieving integrity of the original data. It is better to apply data mining on reduced version of data producing results as the same as or almost the same of original data. Data reduction methods involve data representation, dimensionality reduction and data compression [14]. This paper concentrates mainly on applying CFS and fuzzy rough feature selection.

a) Correlation based Feature Selection (CFS)

Correlation based feature selection (CFS) assesses the value of a subset of attributes by considering the individual predictive ability of every feature alongside the level of redundancy between them. CFS algorithm assembles evaluation formula with specific and reasonable correlation metrics and heuristic search strategy. There are many trials on standard datasets demonstrated that CFS rapidly distinguishes immaterial, repetitive, and noisy features. Also CFS screens relevant features as long as their significance does not emphatically rely on other features. On medical domains, CFS commonly reduced well over a large portion of the features. Much of the time, classification accuracy based on eliminated features gives good accuracy [18], [19].

b) Fuzzy Rough Feature Selection (FRFS)

The rough set attribute selection (RSAR) methodology can just work adequately with datasets containing discrete values. Furthermore, there is no chance to handle noisy data. Because most datasets contain real valued attributes, it is important to perform a discretization step in advance. This is normally actualized by standard fuzzification methods. Fuzzy-rough feature selection (FRFS) gives a method by which discrete or real-valued noisy data can be successfully eliminated without the requirement for user supplied data. Furthermore, this procedure can be implemented with nominal or continuous attributes that can be found in classification and regression datasets [27], [30]-[38].

2) Data Cleaning: To achieve high quality and accuracy of data to any information system data must be cleaned. Data cleaning is defined as the process of discovering and reducing artifacts for improving the data quality that is necessary for building any knowledge discovery and data warehouse [20]. Data cleaning methods differ according to the nature of the problem or area that apply to it but in general used to detect incomplete, inaccurate or unreasonable data and starting to improve such data by correcting what detected.

C. Data Mining and Artificial Intelligence Techniques

- Data Classification

Classification [18] is considered one of the forms of data analysis with supervised learning for extracting models portraying imperative data classes. Such models called classifiers for forecasting discrete or unordered class labels. The process of data classification comprise of two main steps for learning and classification where the model is utilized to anticipate class labels for given or specific data.

- Neural Networks

Neural networks are one of the most popular approaches to machine learning for improving the performance of intelligent systems. Neural network simulate human brain so called biological system that can be used for pattern recognition. Artificial neural networks (ANN) are artificial intelligence techniques used widely to solve pattern recognition and decision support problem in bioinformatics field. Neural network can be combined with different techniques such as producing rough neural network by accumulating neural networks with rough set. [21]

- K-Nearest Neighbours

K-nearest neighbour [22] is a classification technique that accepts the class of an instance to be the same as the class of the closest occurrence to that instance. It receives a similarity metric to quantify the closeness of an instance to others. Nearest neighbour proposes that instances in the data will be independently and indistinguishably distributed, so the instances have the same classification if they are in close proximity. For predicting a class, the algorithm must calculate how far attributes of new and previous differ.

- Naive Bayes

The naive Bayes algorithm utilizes an improved form of Bayes equation to choose which class a novel occurrence belongs to. The back likelihood of every class is calculated. Given the highlight qualities introduced in the occasion, the occurrence is allocated the class with the most elevated likelihood. Equation 1 demonstrates the naive Bayes equation, which makes the presumption that features values are factually autonomous inside every class [19].

$$p(C_i \left| \upsilon_1, \upsilon_2, ..., \upsilon_n \right.) = \frac{p(C_i)\prod_{j=1}^{n} p(\upsilon_j \left| C_i \right.)}{p(\upsilon_1, \upsilon_2, ..., \upsilon_n)} \qquad (1)$$

Where, C_i indicates the class and (v1, v2… v_n) for features values.

- C4.5

C4.5 [19] is algorithms that can represent their training data as a decision tree. C4.5 utilizes a greedy approach that makes use of an information theoretic measure to assemble a decision tree from training data as its guide. Training instances are divided into subsets by selecting an attribute for the root of the tree according to attribute values. C4.5 use gain ratio for ranking and selecting which attribute to be a root for the tree. There are more algorithms that depend upon tree classification such that ID3, NB tree and FT (functional tree) Classifier for 'Functional trees' builds a tree for classification that could have logistic regression capacities at the internal nodes and/or its leaves.

- Decision Table/Naive Bayes Hybrid (DTNB)

Naive Bayes and Decision Tables (DTNB) [23] is hybrid model of combining a simple Bayesian network where the decision table (DT) represents a conditional probability table. Hybrid model learning algorithm (DTNB) continues similarly as the one for stand-alone DTs. The research for each point evaluates and assesses the merit connected with splitting the attributes into two subsets that are disjoint: the first for the DT and the second for NB. DTNB hybrid demonstrated a high performance evaluation compared to applying each in its own.

- K-Star (K*)

K* is an instance-based learner that are means for classifying an instance by instance-based learners. This is made by comparing and contrasting instances with pre-classified samples of pre-defined data sets [24]. Hence, the crucial supposition with comparison states that similar instances have similar classifications. The question here is how to determine that there are two similar instances and how similar they are. Distance function, considered from the corresponding components of instance-based learner, is used to measure instance similarity. The second component of instance-based learner is the classification function. It is responsible for determining the final classification for a new instance produced from instance similarities.

- Sequential Minimal Optimization Algorithm(SMO)

Sequential minimal optimization (SMO) is a simple algorithm that can rapidly tackle the support vector machine-quadratic programming (SVM QP) problem with no additional matrix storage or utilizing numerical QP optimization steps at all. SMO limited the problems overall QP to QP sub-problems by using Osuna's theorem to guarantee convergence [25].

- Data Clustering (K-Means Algorithm)

Data clustering is a task of data mining that group data or objects that have similar properties together to be used to facilitate their processing. Data clustering have applied in many domains such as medical area. There are several clustering algorithms existing in research but k-mean algorithm is popular because of its simplicity in implementation and ability to deliver great results.

The k-means algorithm is an unsupervised mining clustering techniques. K-means is widely applied in bioinformatics and related fields that need to determine the number of clusters that appropriate for specific problem. K-means algorithm includes five steps [14] (Fig. 2):

Step1. Determining number of clusters
Step2. Create clusters randomly and select cluster midpoint
Step3. Assign each object to the group that has the closest centroid
Step4. When all objects has been assigned, re-computation of the positions of new cluster midpoints
Step5. Repeat steps 3 and 4 until a separation is produced between objects that achieved when no longer centroids moved.

Fig. 2. K-means algorithm.

K-means algorithm purpose and goal is minimizing the objective function (squared error function) given by:

$$J(\upsilon) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\left\| x_i - \upsilon_j \right\|)^2 \qquad (2)$$

Where,

'$\|x_i - v_j\|$' is the Euclidean distance between xi and vj. 'C_i' is the number of data points in i th cluster.

'C' is the number of cluster centers.

K-mean clustering method calculates the distance (d) between two objects o_i and o_j by Euclidean separation given as:

$$d(o_i, o_j) = \sqrt{\sum\nolimits_{p=1}^{d} (o_i^p - o_j^p)^2} \qquad (3)$$

The hybrid system proposed in this study evaluates a new pre-processing step. It consists of data cleaning, clustering of data, and data reduction. Moreover, a comparative study between different algorithms for data reduction and the effect of this whole pre-processing step for enhancing classification algorithms is introduced. Data cleaning is important to overcome incomplete or inaccurate data. Data selection, without missing values in columns or rows, participate in achieving more accurate results as pre-processing step. There are many algorithms for clustering process but k-means is accepted for its simplicity and widely used in bioinformatics and related domains.

After data clustering, the new data set is composed of two subsets. The feature reduction step is applied on each subset. Data reduction can be done by many algorithms but only two algorithms are choosen in this step. The first algorithm for data reduction is CFS that rapidly screens immaterial, repetitive, and noisy features. Moreover CFS distinguishes relevant features as long as their significance does not emphatically rely on upon other features.

The core element of CFS is a heuristic [28] that used to evaluate the worth (merit) to specific set of the features. Heuristic used to calculate how a set of feature effect on predicting the label of class through the inter-correlation among them. Heuristic formalization can be displayed in (4):

$$Merit_s = \frac{k\,\overline{r_{cf}}}{\sqrt{k + k\,(k-1)\overline{r_{f\,f}}}} \qquad (4)$$

Where Merits is the heuristic 'merit' of a feature Subset S that contain k features, $\overline{r_{cf}}$ is the mean feature class correlation (f$\in$S) and $\overline{r_{f\,f}}$ is the average feature-feature inter-correlation.

The second algorithm is FRFS that have many advantages when working with discrete, real values, noisy, nominal or continuous terms of data without more user supplied data. The fuzzy rough set has utilized the vagueness of fuzzy sets with rough sets concepts of indiscernibility. FRFS generalizes the rough set by a fuzzification strategy which remains the basic values of attributes unchanged yet produces a collection of fuzzy sets for each one. Fuzzy partitioning of the input space or fuzzy similarity relation for approximating fuzzy concept (5) can be used in the FRFS algorithm implementations.

$$\mu_{R_a}(\chi, y) = \max\Big(\min\Big(\frac{(a(y)-(a(\chi)-\sigma_a))}{(a(\chi)-(a(\chi)-\sigma_a))}, \frac{((a(\chi)+\sigma_a)-a(y))}{((a(\chi)+\sigma_a)-a(\chi))}, 0\Big) \qquad (5)$$

FRQuickReduct [29] implements the FRFS basing on the dependency (6) that calculates the membership dependency degree between the fuzzy attributes and the equivalence classes.

$$\gamma'_p(Q) = \frac{\sum_{\chi\in U}\mu_{POS_{R_p}(Q)}(\chi)}{|U|} \qquad (6)$$

Where $\mu_{POS_{R_p}(Q)}(\chi) = \sup_{\chi\in U/Q} \mu_{R_p x}(x)$ (7)

The FRQuickReduct(C,D) is illustrated in Fig. 3. Where C is the set of all conditional attributes and D is the set of decision attributes.

```
$R = \{\}, \gamma'_{optimal} = 0, \gamma'_{old} = 0$
$do$
$T = R$
$\gamma'_{old} = \gamma'_{optimal}$
$\forall \chi \in (C - R)$
    $if\ \gamma'_{R\cup\{\chi\}}(D) > \gamma'_T(D)$
        $T = R \cup \{\chi\}$
            $\gamma'_{optimal} = \gamma'_T(D)$
    $R = T$
  $until\ \gamma'_{optimal} = \gamma'_{old}$
  $retrieve\ R$
```

Fig. 3. FRQuickReduct algorithm.

For selecting which attributes ought to be appended to the candidate reduct, the algorithm utilizes and employs the dependency function γ′. The stopping criteria is when there are no attributes that increase the dependency. The algorithm finishes and gets the reduct.

The next stage of the system is merging reduced features of different clusters again. This methodology, clustering then reduction and merging the data again, prevents eliminating attributes that can participate with any degree to the classification accuracy.

The last step compares more than one algorithm for classification to test to what extend the pre-processing step affects classification of medical data. This study used neural network (NN), K-nearest neighbours, Fuzzy-rough K-nearest neighbours, Discernibility NN classifier, Naive Bayes ,K-Star(K*), Functional trees, C4.5, decision table/naive Bayes hybrid and training a support vector classifier by sequential minimal optimization algorithm(SMO).The proposed framework that display the main components is represented in Fig. 4.

From Fig. 4 the whole system can be summarized in the following phases:

- Data pre-processing:
  - Noise and missing values handling
  - Data clustering
  - Feature extraction
  - Merging data to produce one dataset
- Applying classification algorithms
- Testing the model

This model includes many comparative studies as follows:

a) The first between CFS and FRFS and their effect on classification algorithms.

b) The second with hybrid model between CFS and FRFS in applying them directly on the whole data set vice versa implementing them on each cluster with its own and also the study of their effect on enhancing the accuracy of classification algorithms.

c) The third among classification algorithms with time complexity for original data and proposed model.

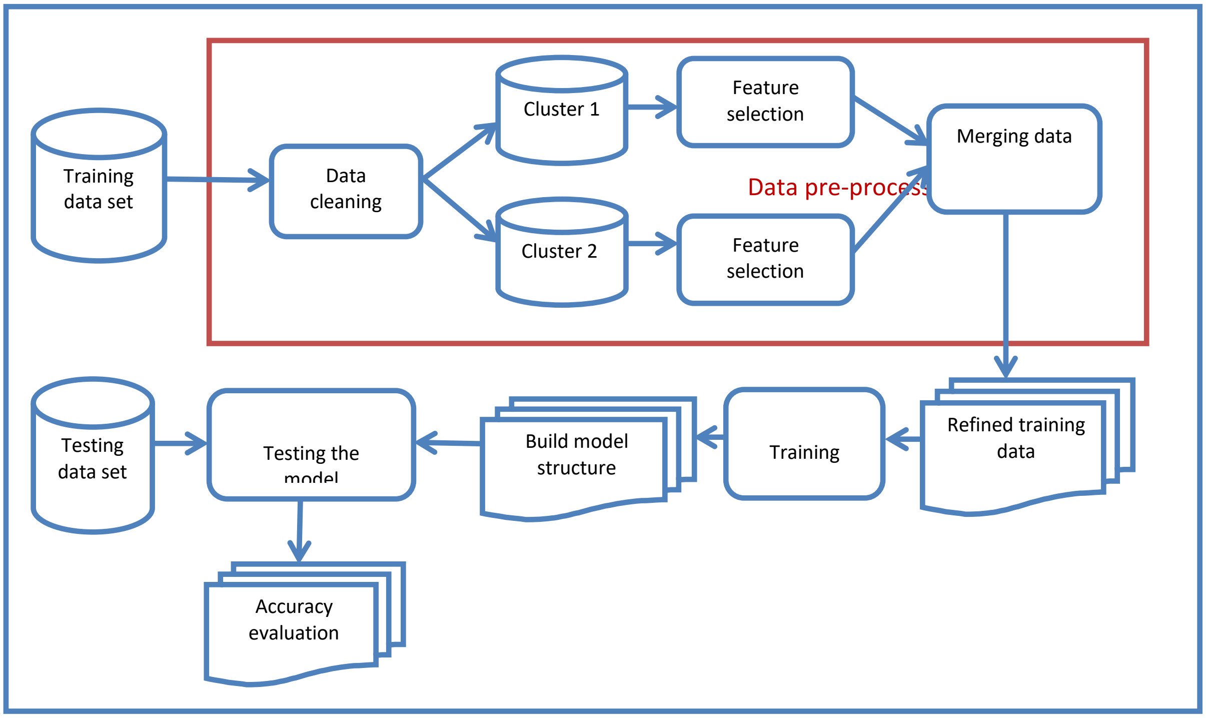

Fig. 4. The proposed framework.

IV. Experiments and Results

a) Data Set

For the examination breast cancer data set from UCI machine learning repository will be utilized to test the model [26]. The highlights of data sets are given in Table I.

The data set contains 699 cases about patients who had experienced surgery for breast cancer. The yield values are either 2 or 4 demonstrating that resting tumor protuberance (benign) or risky bump (malignant). Nine different fields are esteemed from 1 to 10 in addition to ID number, which are itemized in Table II. The undertaking is to figure out whether the identified tumor is benign (2) or malignant (4) given estimations of nine characteristics portrayed in Table II.

TABLE I. Data Set Description

Dataset characteristics	**Multivariate**	**Attributes**	**10**
Attribute characteristics	Integer	Instances	699
Missing values	Yes	Class	2

TABLE II. DATASET ATTRIBUTES

Attribute	**Domain**
Clump Thickness	[1, 10]
Uniformity Cell Size	[1, 10]
Uniformity Cell Shape	[1, 10]
Marginal Adhesion	[1, 10]
Single Epithelial cell Size	[1, 10]
Bare Nuclei	[1, 10]
Bland Chromatin	[1, 10]
Normal Nucleoli	[1, 10]
Mitoses	[1, 10]
Class	{2,4}

From the whole data set there are 458 instances for benign and 241 instance for malignant. The class instead of 2 and 4 we replace them by 0 and 1 for easy processing.

b) Clustering and Reduction

The first step in the proposed model is preparing data set for clustering and handling missing and noisy data. In the process of clustering, K-means clustering algorithm is used for its simplicity. The K attribute value is 2 clusters that suites the nature of data and their classification. Rapid Miner Studio was used for data clustering. After data clustering, there have been 2 sub datasets. One cluster has 354 instances and the other has 345 instances. For each sub data set, data reduction was applied by correlation feature selection and fuzzy rough feature selection algorithms.

The WEKA tool was used to apply the reduction algorithms. Feature reduction showed that the correlation feature selection (CFS) algorithm keeps the same number of attributes as in the original dataset while applying on clustered data yields 8 attributes. The fuzzy rough feature selection (FRFS) algorithm yields the same number of attributes (7 attributes) in both models of reduction (FRFS directly on the original data and the clustered data sets) but with different attributes. From the result of reduction, the level of reduction is not large but the accuracy of medical data classification is the most important factor in patient treatment.

c) Classification Algorithms

At this step of the model, the research tries to investigate different machine learning algorithms for classifying data and testing how pre-processing step of (cleaning +clustering +reduction+ merging) have affected the improvement of classification algorithms accuracies.

Classification algorithms are implemented by WEKA tool on the new data produced from proposed model compared to applying the same algorithms on original and reduced data.

The classification algorithms were tested by 10 folds cross validation and percentage split with 80-20 for training and test respectively. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.

Tables III and IV show the classification algorithms and their results for accuracy metric for the proposed model compared to original data for both feature reduction algorithms. From the results, there are enhancements in the accuracy of the pre-processing with clustering against applying the same algorithms directly on the original data or after reducing the feature directly on the original data by CFS and FRFS.

From Fig. 5 and 6 the proposed model proved that the clustering added to the pre-processing step has a main role in improving the accuracy of classification algorithms.

Fig. 7 shows that there are enhancements in the proposed model for both reduction algorithms with two test modes. Also it is noted that by using FRFS the accuracy levels exceeds the accuracy levels of CFS.

The proposed model increases the efficiency with ratio up to 2.92 than using just reduction techniques. Fig. 8 and 9 demonstrates the levels of improvements for all classification algorithms when using the proposed pre-processing step. It shows that all algorithms of classification change for the better. In the other hand, using feature extraction algorithms directly on data set may decrease or increase some algorithms accuracy which is not an increase rate of the proposed system.

The enhancements of proposed model can be summarized in Fig. 10 in add value property. In addition to improving the accuracy of classification algorithms, the proposed model reduces the time consumption for those algorithms. Time complexity is important factor combined with accuracy in dealing with critical fields of human life. Fig. 11 and 12 show the time complexity (in seconds) for building the classification model under the two reduction algorithms for original data, reduced data and the proposed model.

TABLE III. THE ACCURACY OF CLASSIFICATION BY USING FRF

Classifier	**Original data**		**Original +FRFS**	**Proposed model**	
	10-folds	**Percentage split 80-20**	**Percentage split 80-20**	**10-folds**	**Percentage split 80-20**
FRNN	95.7	95.8	95.8	96.4	98.6
NN	96.4	96.4	96.4	96.9	97.9
SMO	96.7	96.6	96.5	96.9	97
NAIVIBAYES	96.2	95.8	95.8	96.7	96.1
DTMB	97	95.8	96.5	97.3	97.2
J4.8	94.6	93.4	96.4	96	96.6
FT	96.9	95.8	95.8	97.3	96.6
Discernibility NN	96.7	97.2	97.2	97.3	98.9
MLP	95.9	95.7	94.5	96.9	97.2
k-star	81.8	77.7	97	96.1	98.6

TABLE IV. THE ACCURACY OF CLASSIFICATION BY USING CFS

Classifier	**Original data**		**Original +CFS**	**Proposed model**	
	10-folds	**Percentage split 80-20**	**Percentage split 80-20**	**10-folds**	**Percentage split 80-20**
FRNN	95.7	95.8	95.1	96	96.5
NN	96.4	96.4	97.2	96.9	97.9
SMO	96.7	96.6	96.5	96.7	96.8
NAIVIBAYES	96.2	95.8	95.8	96.2	96
DTMB	97	95.8	95.8	97.2	96.9
J4.8	94.6	93.4	93.4	95.2	95.4
FT	96.9	95.8	95.8	96.9	96.5
Discernibility NN	96.7	97.2	97	97.2	98
MLP	95.9	95.7	95.8	96.4	96
k-star	81.8	77.7	97	96	98.6

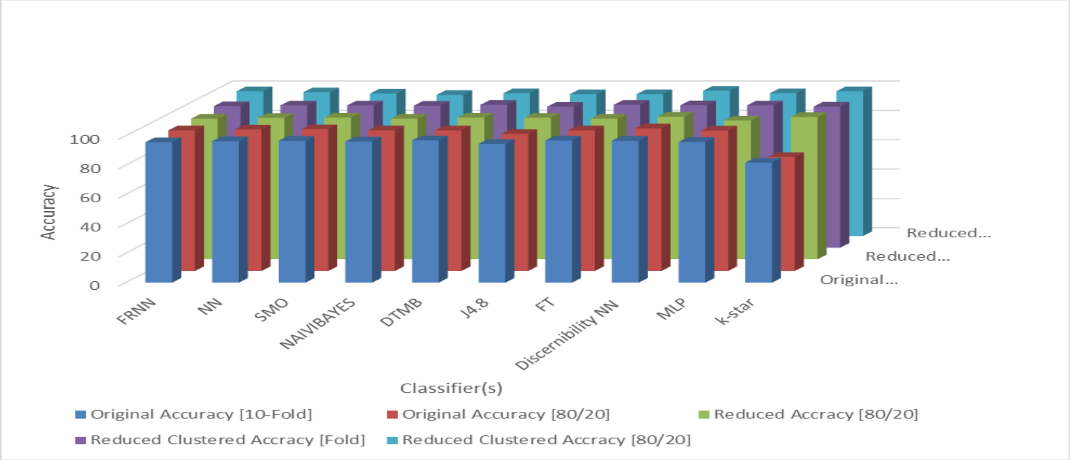

Fig. 5. Classification accuracies using FRFS.

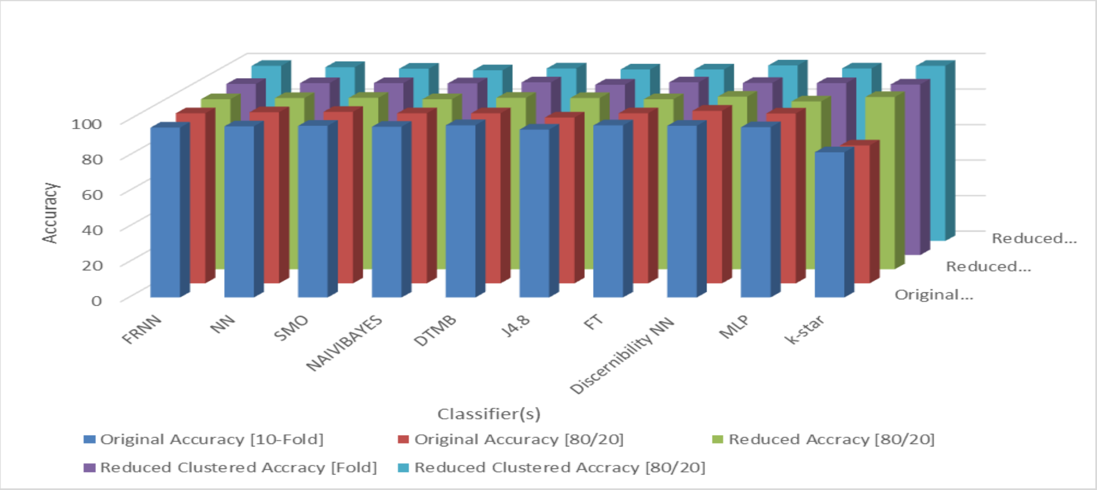

Fig. 6. Classification accuracies using CFS.

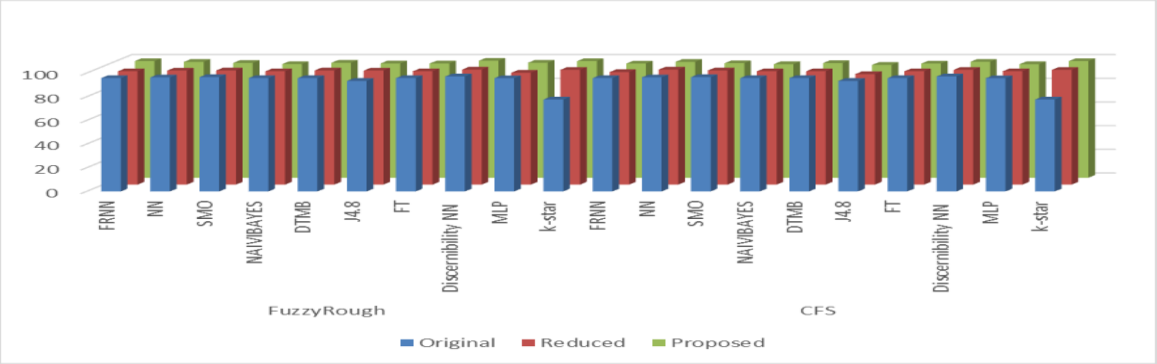

Fig. 7. Classification algorithms for CFS and FRFS.

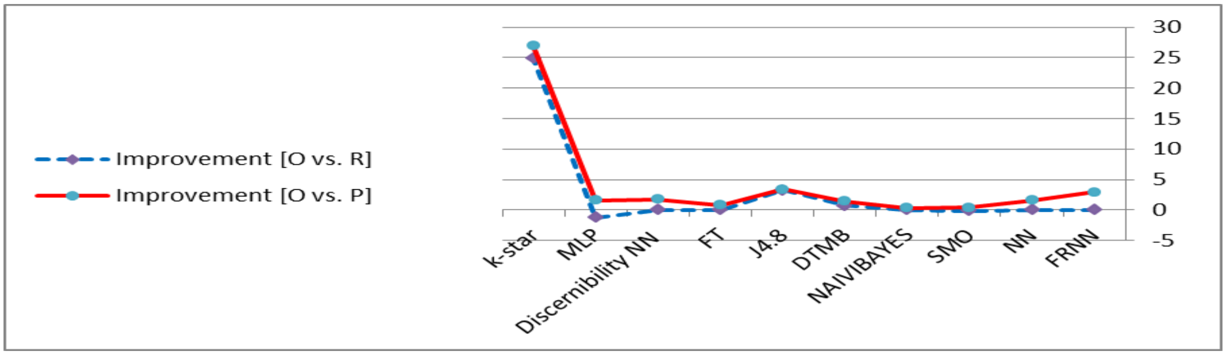

Fig. 8. The level of improvements of reduced (R) over original (O) against proposed (P) over original (O) (FRFS).

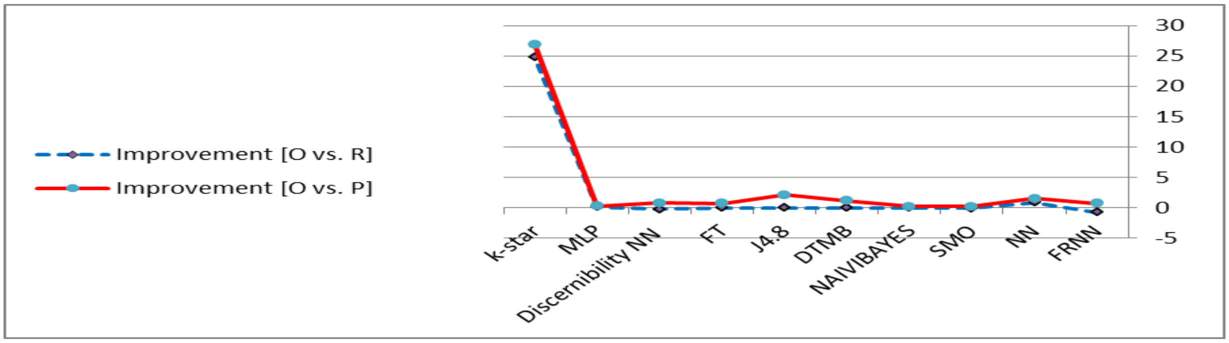

Fig. 9. The level of improvements of reduced (R) over original (O) against proposed (P) over original (O) (CFS).

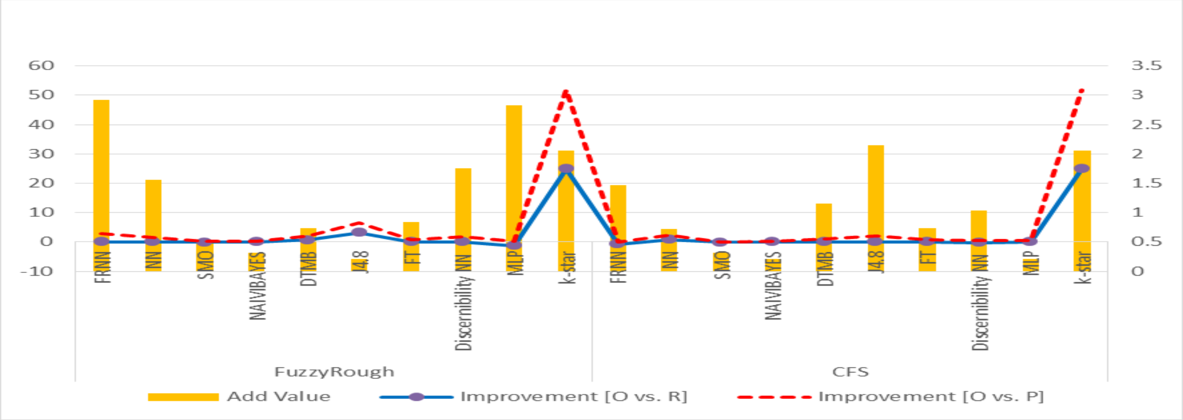

Fig. 10. Add value for proposed model (for FRFS and CFS).

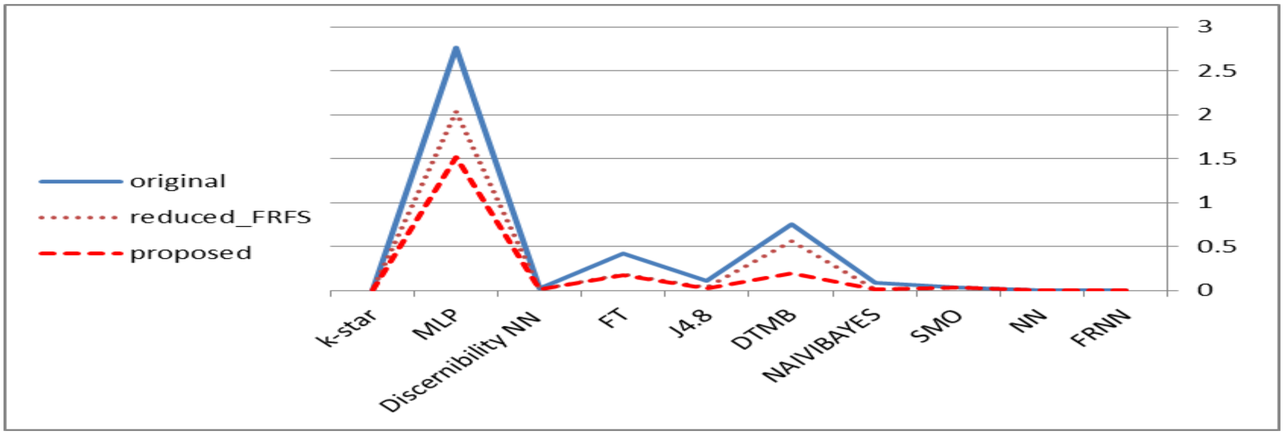

Fig. 11. Time for building model for original data, reduced and proposed model using FRFS.

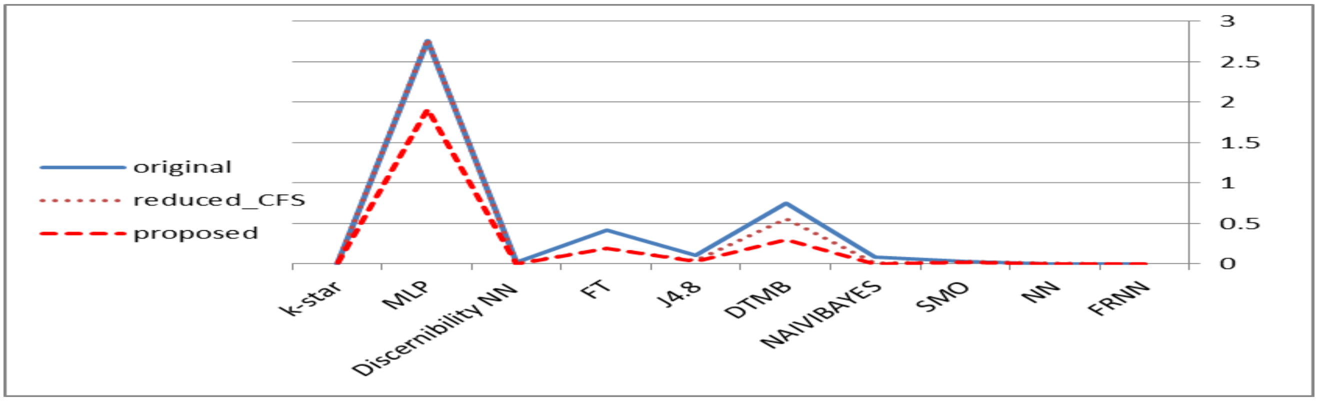

Fig. 12. Time for building model for original data, reduced and proposed model using CFS.

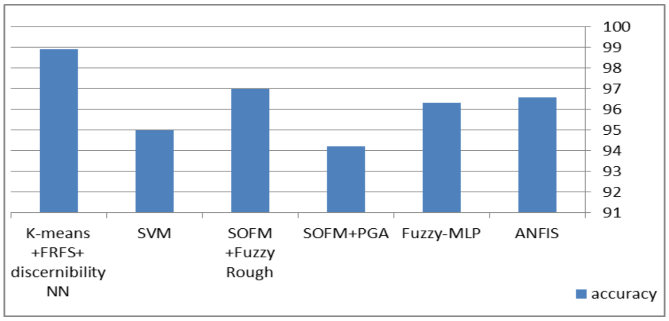

Fig. 13. Proposed system vs. previous studies.

The proposed model showed not only enhancements in classification algorithms but suggest a new hybrid models that can be compared to previous studies on the same data set. The proposed system reaches in classifying breast cancer with accuracy to (98.9%) with a hybrid composed of (FRFS +K-means +Discernibility NN) and Table V shows this result with other results in previous studies. The relation between the proposed system and previous studies can be graphically displayed in Fig. 13.

TABLE V. COMPARISON BETWEEN PROPOSED AND PREVIOUS STUDIES ON BREAST CANCER DATA SET

Classification models	Accuracy
ANFIS	96.59
Fuzzy-MLP	96.3
SVM	95
SOFM+PGA	94.2
SOFM +Fuzzy Rough	97
K-means +FRFS+ discernibility NN	98.9

V. CONCLUSION

Medical data provides a challenging field for data mining researchers. Machine learning algorithms were used to mine information from ambiguous and vague concrete data. Data pre-processing intends to enhance the final accuracy of medical data classification. A hybrid pre-processing model to enhance the performance of classification algorithms has been presented. Such model combines K-means clustering algorithm with fuzzy rough feature reduction or correlation feature reduction to achieve effective data reduction. The proposed model has been applied on breast cancer data set from UCI machine learning repository. Simulation results have shown the effectiveness of the proposed model in enhancing the performance of classification algorithms. Furthermore it has been proven that fuzzy rough feature selection is better than correlation feature selection in data reduction, in addition, it increases the accuracy of classification. Compared to previous studies on the same data, it has been shown that the hybrid model of k-means, fuzzy rough feature selection and discernibility nearest neighbour is more efficient than other algorithms in the same field.

REFERENCES

[1] Nikita Jain, Vishal Srivastava. "Data mining techniques: A survey paper." IJRET: International Journal of Research in Engineering and Technology, vol.2,no.11, (2013): pp. 2319-1163.

[2] Data Mining Software. "Www.Chel.Com.Ru/~Rav/Data_Mining_Software.Html". N.p., 2015. Web. 1 May. 2015.

[3] Joseph, Zernik."Data Mining As a Civic Duty – Online Public Prisoners Registration Systems". International Journal on Social Media: Monitoring, Measurement, Mining vol.-1, no.-1 (2010): pp. 84-96.

[4] S.S.Baskar , Dr. L. Arockiam ,S.Charles ." A Systematic Approach on Data Pre-processing In Data Mining" COMPUSOFT : International Journal of Advanced Computer Technology vol.2 (2013):pp.335-339.

[5] Koh, Hian Chye, and Gerald Tan. "Data mining applications in healthcare."Journal of healthcare information management vol.19, no.2 (2011):pp. 56.

[6] Hersh, W., Jacko, J. A., Greenes, R., Tan, J., Janies, D., Embi, P. J., & Payne, P. R." Health-care hit or miss? "Macmillan Publishers Limited, vol.470 (2011):pp327-329.

[7] Yukti Agarwal, Hari Mohan, Pandey, "Performance Evaluation of Different Techniques in the Context of Data Mining- A Case of an Eye Disease" , 5th International Conference- Confluence The Next Generation Information Technology Summit (Confluence)- IEEE, (2015):pp. 72-76.

[8] ANUSHYA, A.; PETHALAKSHMI, A,. "A comparative study of fuzzy classifiers on heart data", 4th International Conference on Computer and Communication Technology (ICCCT) on. IEEE, (2013): pp 17-21.

[9] Kavita Rawat, Kavita Burse. "A Soft Computing Genetic-Neuro fuzzy Approach for Data Mining and Its Application to Medical Diagnosis" ,International Journal of Engineering and Advanced Technology (IJEAT) Vol.3 (2013) :pp.409-411

[10] SHUKLA, Manish; AGARWAL, Sonali." Hybrid approach for tuberculosis data classification using optimal centroid selection based clustering". In: Engineering and Systems (SCES), Students Conference on. IEEE, (2014): pp. 1-5.

[11] Hazlina Hamdan and Jonathan M. Garibaldi, "A Framework for Automatic Modelling of Survival using Fuzzy Inference" , WCCI IEEE World Congress on Computational Intelligence, Brisbane, Australia , (2012): pp.1-8.

[12] MARCANO-CEDEÑO,Alexis;QUINTANILLA-DOMÍNGUEZ,Joel; ANDINA, Diego." WBCD breast cancer database classification applying artificial metaplasticity neural network". Expert Systems with Applications, vol.38, no.8, (2011):pp. 9573-9579.

[13] Jahanvi Joshi, Rinal Doshi, Jigar Patel. "Diagnosis of Breast Cancer using Clustering Data Mining Approach" ,International Journal of Computer Applications ,vol.101, No.10, (2014), pp.13-17.

[14] Han, J., M. Kamber, and J. Pei. "Data Mining, Second Edition: Concepts and Techniques ".The Morgan Kaufmann Series in Data Management Systems. ISBN-13: 978-1558609013. (2006).

[15] Myatt, Glenn J. Making sense of data: "A practical guide to exploratory data analysis and data mining". John Wiley & Sons, ISBN-13: 978-0-470-07471-8, (2007).

[16] Alberto Guillén,Amaury Lendasse, Guilherme Barreto," Data Preprocessing and Model Design for Medicine Problems", Computational and Mathematical Methods in Medicine ,vol.2013(2013):pp.1.

[17] Salleb, Ansaf , Christel Vrain ."An Application Of Assosiation Knowledge Discovery And Data Mining". Springer Verlag. LNAI 1910 (2000): pp. 613-618.

[18] Hall, Mark A. "Correlation-based feature selection for machine learning". Diss. The University of Waikato, 1999.

[19] Lu, Xinguo, et al. "A novel feature selection method based on CFS in cancer recognition." Systems Biology (ISB), IEEE 6th International Conference on. IEEE, (2012):pp.226-231

[20] Kumar, R. Kavitha, and R. M. Chadrasekaran. "Attribute correction–data cleaning using association rule and clustering methods." International Journal of Data Mining & Knowledge Management Process (IJDKP) vol.1, no.2 (2011): pp. 22-32.

[21] R. Andrews, J. Diederich, A. B. Tickle, "A survey and critique of techniques for extracting rules from trained artificial neural networks", Knowledge-Based Systems,vol.- 8,no.-6, (1995): pp.-378-389.

[22] Syed, Muhammad. "Attribute weighting in k-nearest neighbor classification." (2014).

[23] HALL, Mark; FRANK, Eibe." Combining Naive Bayes and Decision Tables". In:FLAIRS Conference. (2008): pp. 318-319.

[24] CLEARY, John G., et al. "K*: An instance-based learner using an entropic distance measure". In: Proceedings of the 12th International Conference on Machine learning. (1995): pp. 108-114.

[25] PLATT, John, et al. "Sequential minimal optimization: A fast algorithm for training support vector machines". Microsoft Research MST-TR-98-14 (1998):pp.1-21.

[26] Ics.uci.edu, (2015). Donald Bren School of Information and Computer Sciences @ University of California, Irvine. [online] Available at: http://www.ics.uci.edu [Accessed 1 May 2015].

[27] Mona Gamal,Ahmed Abou El-Fetouh, Shereef Barakat ."A Fuzzy Rough Rule Based System Enhanced By Fuzzy Cellular Automata". (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 4, no.5, (2013):pp.1-11.

[28] Hall, Mark A., and Lloyd A. Smith. "Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper." In FLAIRS conference, (1999):pp. 235-239.

[29] Jensen, Richard, and Qiang Shen. "New approaches to fuzzy-rough feature selection." Fuzzy Systems, IEEE Transactions on VOL.17, No.4 (2009):pp. 824-838.

[30] Ibrahim M. El-Hasnony, Hazem El Bakry, and Ahmed A. Saleh, "Comparative Study among Data Reduction Techniques over

Classification Accuracy," International Journal of Computer Applications, vol. 122, No. 2, (July 2015): pp. 8-15.

[31] Ibrahim M. El-Hasnony, Hazem M. El Bakry, and Ahmed A. Saleh, "Classification of Breast Cancer Using Softcomputing Techniques," International Journal of Electronics and Information Engineering, vol. 4, No. 1, (Jan. 2016): pp. 45-53.

[32] Ibrahim M. El-Hasnony, Hazem M. El Bakry, Ahmed A. Saleh, "Data Mining Techniques for Medical Applications: A Survey," Proceedings of WSEAS 16th International Conference on Mathematical Methods, Computational Techniques and Intelligent Systems (MAMECTIS '14), Lisbon, Portugal, (October 31- November 1, 2014): pp. 205-212.

[33] Hazem M. El-Bakry and Nikos Mastorakis, "Fast Detection of H1N1 and H1N5 Viruses in DNA Sequence by using High Speed Time Delay Neural Networks," International Journal of Computer Science and Information Security, vol. 9, no. 11, (November, 2011): pp. 101-108.

[34] Ahmed A. Radwan, Hazem M. El-Bakry, and Hager El Hadaad, "A New Expert System for Pediatric Respiratory Diseases by Using Neural Networks," Proc. of the 2nd International conference on Applied Informatics and Computing Theory, Prague, Czech Republic, (September 26-28, 2011) : pp. 296-306.

[35] A. A. Radwan, H. M. El-Bakry and H. M. El-hadaad, "A New Expert System For Pediatric Respiratory Diseases By Using Neural Network," International Journal of Computer Science and Information Security, vol. 9, no. 8, (August, 2011): pp. 11-17.

[36] Hazem M. El-Bakry, "Fast Virus Detection by using High Speed Time Delay Neural Networks," Journal of Computer Virology, vol.6, no.2, 2010, pp.115-122.

[37] Hazem M. El-Bakry, "A Novel High Speed Neural Model for Fast Pattern Recognition," Soft Computing Journal, vol. 14, no. 6, 2010, pp. 647-666.

[38] Samir M. Abd El-Razek, Waeil F. Abd El-Wahed and, Hazem M. El-Bakry, "MUVES: A Virtual Environment System for Medical Case Based Learning," International Journal of Computer Science and Network Security, vol. 10, no. 9, September 2010, pp. 159-163.