EasyChair Preprint

№ 59

Hybrid Multistage Fuzzy Clustering System for Medical Data Classification

Maryam Abdullah, Fawaz S. Al-Anzi and Salah Al-Sharhan

April 10, 2018

Hybrid Multistage Fuzzy Clustering System for Medical Data Classification

Maryam Abdullah
Computer Engineering Department
Kuwait University, Kuwait
mar.alkandari@gmail.com

Fawaz S. Al-Anzi
Computer Engineering Deparmtnet
Kuwait University, Kuwait
fawaz.alanzi@ku.edu.kw

Salah Al-Sharhan
Gulf University for Science and Technology, Kuwait
AlSharhanS@gust.edu.kw

***Abstract*— Due to the rapid development in technology nowadays, massive amount of data are available. In medicine, decision making is entirely based on the hidden information in these massive data. For that reason, data mining and machine learning technologies provide powerful tools for knowledge discovery within data. Two main techniques are used interchangeably: clustering and classification. In machine learning, clustering is an unsupervised learning technique while classification is a supervised learning method. These techniques are capable of extracting useful patterns and information which aid the process of data analysis and clinical decisions. This research presents a recent study of these techniques in the medical field during the past five years. Moreover, this paper proposes a hybrid multistage fuzzy clustering system applied to medical data classification. In the proposed system, two fuzzy clustering algorithms specifically FCM and GK were initially employed to obtain the membership values. These weights are then used in the second stage of the system as additional informative features to improve the classification process completed by SVM algorithm. Wisconsin Breast Cancer dataset, real-world application, obtained from UCI were used in the experiments. The results of the experiments show that the additional weights further improve the classification accuracy with 99.06% and 100% sensitivity.**

Keywords*— machine learning; supervised learning; unsupervised learning; clustering; classification; SVM; WBC

I. INTRODUCTION

As a result of the scientific revolution, machine learning has evolved out of artificial intelligence, AI. In the early times of AI, scholars were interested in developing machines that are able to learn from data. They have developed several methods to approach this problem. However, later, machine learning was recognized as a separate field. It began to flourish during the 1990s. The field shifted its goal to tackle solvable problems of practical nature rather than achieving artificial intelligence. Machine learning changed its approach towards borrowing techniques and models from statistics and probabilities. [1]

Machine learning and data mining most often employ the same methods and they overlap significantly, yet while machine learning emphases on prediction, based on known properties learned from the training data, data mining concentrates on the discovery of prior unknown properties within the data. Data mining utilizes several machine learning methods, but for different purposes; on the other hand, machine learning as well employs data mining methods as unsupervised learning or as a preprocessing phase to enhance the learning performance.

Machine learning algorithms were designed from the very beginning and employed to analyze medical data sets. Nowadays, machine learning provides several crucial tools for intelligent data analysis. This research concerns only about supervised and unsupervised learning algorithms. These techniques were heavily employed in medical field. Medical area has huge amount of data that require processing and analysis in order to extract useful information that sometimes might save a human life. Medical data include patient records, test results, or some type of images such as X-rays, MRI and CT scans. In order to analyze these data, supervised and unsupervised learning techniques are necessary to facilitate data handling and decision making.

For more efficiency and usefulness in solving medical diagnostic tasks, a machine learning algorithm must have the following desired features. The algorithm must achieve a good performance. This occurs when a technique is able to successfully handle missing and noisy data. Also, it must has the ability to illustrate decisions and reduce the number of necessary tests taken to obtain reliable diagnosis. Moreover, the technique should be able to extract significant information from the existing data with the ability to diagnose new cases accurately. [2]

In this following sections, the two major types of machine learning techniques, the concepts behind them and some examples of these methods are also highlighted; in addition to their application in medical area. Implementation methodology section which introduces the proposed system in details, performance evaluation, and experimental results are then presented. Finally, this paper concludes the work and suggest future works.

II. MACHINE LEARNING

In machine learning, there are three major learning styles for an algorithm: supervised learning, unsupervised learning, and reinforcement learning, represented in figure 1. Supervised learning algorithm means learning with a teacher; it uses previously defined labels in order to construct a general model that is able to map input to output. While unsupervised learning finds hidden structure in data that has no labels. Reinforcement learning is based on rewards and punishments; the training data is only provided as feedback to the program's actions in a

dynamic environment; this includes applications like driving a vehicle or playing a game against an opponent.

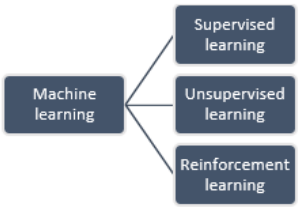

Figure 1: Machine learning styles

A. *Supervised Learning*

Supervised learning is one of the most popular task frequently used in intelligent systems where large number of supervised, classification in data mining, techniques were developed earlier. In this type of learning, the labels must be predefined prior in order to be able to match them to either classes. Supervised learning algorithms starts first by analyzing the training data and then continue to generate a classifier which predicts the output of new valid input. The learning algorithm requires generalization from the training data to new unseen objects. There are several steps that should be completed to solve a given supervised learning problem listed below:

1. Identify the type of training examples.
2. Collect the training set.
3. Identify the feature representation of the input of learned function.
4. Identify the learning function structure and corresponding learning technique.
5. Complete the algorithm design.
6. Evaluate the performance of the learning function.

In the first step the researcher decides the type of data to be used. For example, the type might be a single character, or a string. While in step 2, the training set should represent a real-world practice of the function. Thus, the set of input and its corresponding output are collected either from measurements or could be from human experts. In step 3, the performance, accuracy, of the learning function is strongly influenced by the way the input object is represented. All the objects in the input are transformed into a vector of features. This vector includes descriptive number of features for an object. Researchers must avoid large numbers of features in order to avoid curse of dimensionality. The number of features should be reasonable and informative enough in order to achieve accurate prediction of the output. In step 4, several techniques are available for learning function such as decision trees and support vector machine algorithms. In the fifth step, execution of the algorithm is completed on the training set gathered earlier. Some of them need a predefined parameters determined by the user. These parameters could be adjustable if necessary via performance optimization performed on a subset, named as validation set, of the training set, or by using cross validation. Finally, in the last step, the accuracy of the learning function is evaluated and then measured after adjusting the required parameters on a testing set that differs than the training set. [3]

Supervised learning techniques or (classification algorithms) differ based on the learning techniques used within the method. These include perceptron based learning similar to Neural Networks [4] , instance based learning like K Nearest Neighbor [5], logic based learning such as Decision Trees [6], statistical learning similar to Support Vector Machine [7]. Moreover, might scholars combine some of these algorithm which forms a type of systems called ensemble systems; thus to improve the classification accuracy [8]. However, this research mostly concerns about statistical learning algorithms, in particular SVM.

B. *Unsupervised Learning*

In unsupervised learning, clustering is considered the most essential question in machine learning. There are no clear definitions for clustering yet the agreement came to describe a classic ones as follows [9]:

1. Instances within the same cluster must be as similar as possible.
2. Instances of different clusters must be as different as possible.
3. Measurement of similarities and differences must be clearly stated and should be practically sounds.

There aren't well definition to what factors could construct a cluster leaving many applications with an overlapping clusters where objects are not well separated. However, this might lead to some issues because most clustering of the problems require well separated groups where there are no overlapping between clusters, in other words, they seek crisp classification. To overcome this problem, fuzzy clustering introduces partial belonging of objects to other clusters. Some examples of why we need clustering can appear in grouping documents that are related to facilitate browsing, or to obtain proteins and genes that share similar functionality, or to provide a clusters of spatial locations disposed to earthquakes. Yet there are other reasons for using clustering techniques. One reason is to efficiently find the nearest neighbor to a particular point. Clustering has been used in wide range of applications such as pattern recognition, psychology, biology, statistics, information retrieval, and medical diagnosis [10]. Examples on Fuzzy clustering techniques include Fuzzy C-mean (FCM) [11], Mountain method (MM) [12], Gustafson Kessel (GK) [13], and Fuzzy C-Shell clustering algorithms [14].

There are number of steps that must be carried out during clustering in order to obtain successful and accurate clusters. [9] The steps that describe the standard process of clustering are shown in Figure 2.

Figure 2: Process of clustering

C. *Medical Applications*

Classification techniques were previously employed for various purpose within medical area such as disease diagnosis,

image segmentation and cancer detection. In this section, a review of implementation of these algorithm in medical field during the past few years is presented.

In [15], a novel combination of five heterogeneous classifiers is presented. To determine final predictions, weighted voting method were employed. Four dissimilar breast cancer data sets were used in the experiments for performance evaluation. For further enhancement, feature selection, pre-processing technique, were applied. Comparisons between the proposed method and other works were performed and the results favor the proposed framework with 97.42% classification accuracy. In [16], a new weighted Naïve Bayes classifier is presented for breast cancer detection. Several experiments were completed to evaluate the performance of the proposed technique. A five-fold cross validation test were employed in the experiments. WBC dataset were used with removal of missing values, around 16 instances. Different metric measures were used for performance evaluation namely sensitivity, specificity and accuracy. The results were compared to other existing work and show that the proposed algorithm is better than the standard Naïve Bayes and outperform other works with 98.54% accuracy rate. In [17], ensemble trees classifier were used for Breast cancer classification. The hybrid approach includes CART classifier with feature selection as well as bagging technique. The experiments were conducted using three breast cancer datasets. Feature selection method was employed in combination with CART to eliminate least significant attributes. The results examined with bagging were compared with the one without bagging. The results of the hybrid technique including feature selection with CART and bagging showed an improved accuracy. In [18], a comparison study has been presented among different classifiers including decision tree (J48), Multi-layer Perceptron (MLP), Naïve Bayes (NB), Sequential Minimal Optimization (SMO), and Instance Based for K-nearest neighbor (IBK). The study was completed on different databases of Breast Cancer: Wisconsin breast cancer (WBC), Wisconsin diagnostic breast cancer (WDBC) via using classification accuracy and confusion matrix depending on 10-fold cross validation method. Moreover, fusion at classification level between these classifiers was also presented for every dataset. Several experiments were completed to investigate each classifier performance. In [19], an empirical comparison is completed using various supervised learning algorithms. The paper presents a study on performance criterion of different machine learning tools such as SVM, NB, RBF networks, Decision Tree (J48) and simple CART for disease detection. Various datasets were used in this study: binary and multiclass. These include Pima Indians diabetes WBC, WDBC and breast cancer tissue obtained from UCI depositary. No pre-processing techniques were used in the experiments. The results show that SVM-kernel is superior and outperform all other classifiers for all datasets in terms of accuracy, sensitivity, specificity and precision.

WBC were used in several experiments in other works. In [20], CART was employed with feature selection (Chi-Squared) for breast cancer classification and the accuracy achieved is about 94.56% while in [21], a comparison of multiple classifiers were completed including C4.5, NN, SVM, and KNN. The results show that SVM achieved best accuracy with 96.99%.

In [22], Bayesian Network is employed to integrate extracted features from three types of brain. Leave-one-out analysis was employed to evaluate the grading performance. The results achieve 92.86% overall accuracy. This results show a promising technique for feature combination of different MRI modalities. In [23], a vision based approach is presented for Parkinson Patients movement analysis. Several comparisons of several classifiers were performed as a classification system is required to assist diagnosis and then treatment. These include RBF-Kernel SVM, KNN, MLP and Radial Basis, RB. Dimensionality reduction technique were used prior experiments for further enhancements. The achieved results was based on the new created 2-D features data. While in general RBF-Kernel SVM and MLP obtained higher accuracy than KNN and RB. In [24], several techniques were used for Diabetes-Mellitus diagnosis. These techniques include Artificial Neural Netowrk, K-fold cross validation and classification, K-nearest Neighbor, and Support Vector Machine. Additional methods involve LDA-SVM and feed forward NN as well as statistical normalization and back propagation.

III. SUPPORT VECTOR MACHINE

Support Vector Machine, SVM, is one type of supervised machine learning technique or a classification algorithm. Support Vector Machines, SVM, are set of supervised techniques that are used for different purposes such as classification, regression and outlier detection. These machines are a subset of a generalized family of linear classifiers. The first support vector machine algorithm was developed by [7] and was called support vector networks. The main idea behind support vector networks is mapping input vectors into feature space Z of a high dimension via some non-linear mapping techniques specified earlier where a linear decision surface, called hyperplane, is created within this space to separate the classes. This hyperplane has unique properties in which it ensures that the network has high generalization ability. Figure 3 below show an example of the steps followed in SVMs for a binary classification problem.

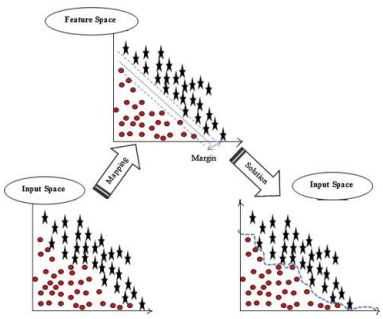

Figure 3: Binary Classification using SVM (Tomar & agrawl, 2015)

It was proven earlier that if an optimal hyperplane is capable of separating the training vectors with no errors, the probability expected value of committing an error on a test sample is bounded to the ratio between the number of training vectors and the expected value of the number of support vectors. The probability equation is shown below:

$$E\,[PR\,(error)] \leq \mathrm{E}\,[\#\,\mathrm{SV}]\,/\,\#\mathrm{TV} \qquad (4)$$

Where SV represents Support Vectors and TV refers to Training Vectors. It should be noted that this bound does not include explicitly the dimensionality of the separation space. It follows that if a small number of support vectors could construct an optimal hyperplane from training dataset thus would result in high generalization, also for infinite dimensional space.

SVM have been used in several application such as face analysis, pattern recognition, and disease prediction. However, SVM were initially developed to solve classification problems, yet recently, it has been used widely for regression applications. [25]

IV. IMPLEMENTATION METHODOLOGY

In this research, hybrid multistage system, shown in Figure 4, is proposed for WBC classification problem. Two main stages are involved in the proposed system. In the first stage, two fuzzy techniques were initially employed for cancer research specifically for breast cancer data, Wisconsin Breast Cancer, by validation against actual cancer. The dataset is a well-known two-class real problem obtained from UCI repository. The two fuzzy clustering algorithms: Fuzzy C-means, FCM, and Gustafson Kesssel, GK, were used at the beginning to obtain the weights of data instances to which of the two clusters they belong. This output provides significant and informative outcome that can be used in the second stage.

A. Clustering Phase

This stage were completed earlier in [26]. The experimental results have shown that a better performance was obtained by FCM over GK with 95% classification accuracy for the former and 91% achieved by the latter. This outcome demonstrates that FCM is more suitable for this particular dataset. In addition, based on these results one more assumption can be made regarding the data distribution. Since FCM searches for spherical clusters, the results achieved by FCM indicate that the data could have Gaussian normal distribution. This outcome leads to the conclusion that some of the fuzzy clustering methods are found to fit some cancer data than other techniques. Since the data has Gaussian distribution, SVM machine with a linear kernel function is preferable for the second stage. In the second stage, the weights resulted in the earlier phase are added to the data as additional informative features. This is expected to result in better performance.

B. Classification Phase

For the second stage of the system, in continuation and based on the first stage, support vector machine is used on the same dataset. By adding the additional outcomes obtained from the fuzzy classifiers, SVM was trained and tested on three data sizes. The datasets were divided into several ratios 50%-50%, 60%-40%, 70%-30% and 80%-20%, respectively. Linear kernel function where used for SVM in the experiments.

α +

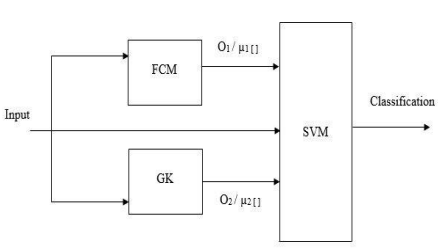

Figure 4: The Architecture of the proposed system

C. Dataset

The UCI Wisconsin Breast cancer (WBC) dataset, a well understood two-class data, is used. The WBC dataset include 699 instances, each of them consists of 11 attributes as follows:

- Sample code number (ID)
- Clump Thickness (CT)
- Uniformity of Cell Size (CS)
- Uniformity of Cell Shape (CSH)
- Marginal Adhesion (MA)
- Single Epithelial Cell Size (EC)
- Bare Nuclei (BN)
- Bland Chromatin (BC)
- Normal Nucleoli (NN)
- Mitoses (MT)

All attributes are real values range between 1 and 10. The class label can have two values; either 2 as benign or 4 as malignant. The first attribute (Sample Code Number) was excluded. There were 16 missing attribute values that are replaced with the mean of its corresponding instance. Following, the data were normalized to unity using min-max normalization technique before being fed to the algorithms.

D. Overall Procedure

The overall procedure of the proposed hybrid system is illustrated in Figure 4. The procedure starts first with employing fuzzy classifiers to obtain the weights of every instance in the dataset. These weights are then used as additional features to train SVM classier and extract the training model, finds the support vectors and the alpha values. The data is randomly divided into three different ratios and the algorithm was trained and tested on all of them.

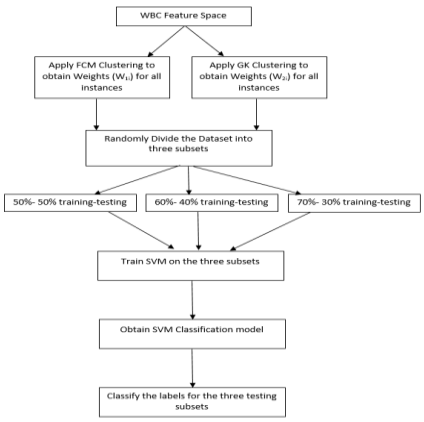

Figure 5: The overall procedure of the proposed system

V. EXPERIMENTAL EVALUATION AND RESULTS

The experiments were completed on windows 7, (64-bit) operating system with 6 GB RAM memory, run on Intel i7-2.40 GHz processor. All algorithms were coded in Matlab. First each fuzzy classifier was trained and tested individually against a real world data then the performance of each one is measured by obtaining the classification accuracy, sensitivity specificity and error of each class, 2 and 4, after that the overall classification accuracy were computed. The classification accuracy of every classifier was calculated using the four performance measures True positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Results of the fuzzy clustering phase can be found in [26].

For the final stage of the proposed hybrid system, the SVM algorithm employed a linear kernel function and the generated variables are as follows:

1. The number of Support Vectors is 50.
2. Alpha is compulsory equal to the number of SV which is 50.
3. Bias value is 0.0676.

In this stage, Support Vector Machin algorithm was employed for further enhancements through addition of the weights obtained in the previous stage to the input of SVM in the classification phase. Thus improves the performance of SVM such that every input instance contributes in the learning of decision surface. This step reduces the impact of outliers and noise, if any, in the data points. SVM was first trained against WBC dataset to extract the classification model that will be used later for testing on different data sizes. The results of the various data sizes ranges between 98% -99%. The best results achieved by the ratio 70%-30% dataset with 99.04% classification accuracy, 100% sensitivity, 98.77 % specificity and finally 0.0096% error, shown in Figure 6, 7, 8, and 9. This means the experiments carried out on the different data sizes have shown the effectiveness of the proposed system against WBC classification problem.

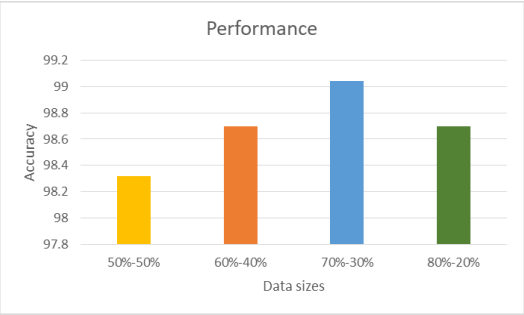

Figure 6: The performance of SVM on different data sizes

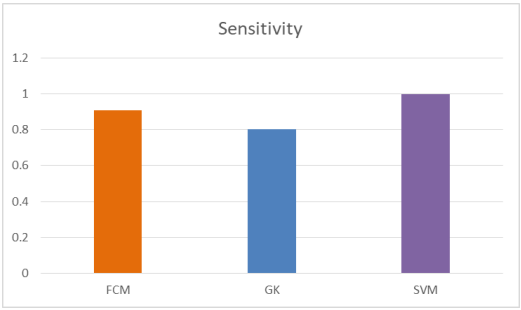

Figure 7: Sensitivity of the whole system

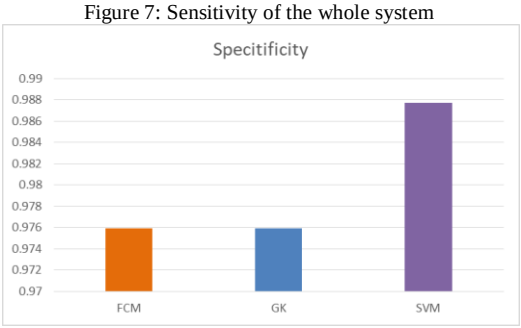

Figure 8: Specificity of the whole system

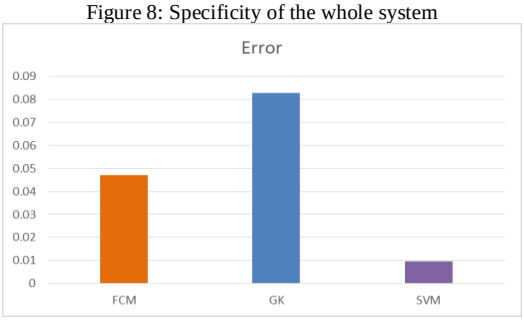

Figure 9: Error rate of the whole system

Moreover, the proposed system were compared to other related work existing in the literature. The Comparison shows that the proposed system have achieved the best performance over the other methods with 99.04 % accuracy, see table 1 and Figure 10 below.

Table 1: Comparison with related work

Reference	Method	Accuracy
[19]	SVM-RBF kernel	96.84
[21]	CART with feature selection (Chi-Sqr)	94.56
[19]	RBF-networks	96.66
[16]	Weighted Naïve Bayes	98.54
[19]	Trees J48	94.59
[17]	Ensemble of DT	97.85
[18]	MLP	95.27
[18]	SMO	96.99
[18]	IBK	94.56
[18]	Fusion -Hybrid	97.28
Proposed System	Hybrid	99.04%

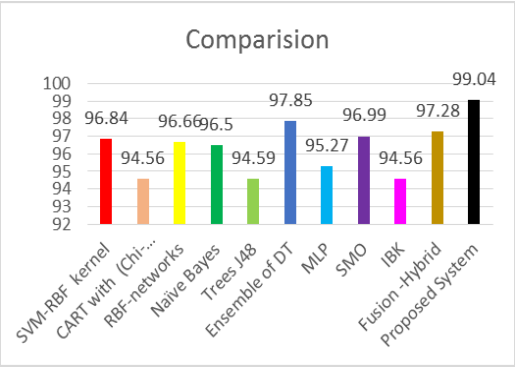

Figure 10: Performance of proposed system against related work

VI. Conclusion and Future Work

This research introduces the essential needs of Supervised and Unsupervised learning for data classification. Moreover, in this paper, a recent study of classification techniques in medical area over the past five years is presented. The paper proposed a hybrid multistage system based on fuzzy clustering for medical data classification. The system involve two main stages. In the first phase fuzzy clustering is employed to generate the weights of every instance in the dataset to which class it belongs to introduce additional significant features added to the data. The data is then fed to SVM in the second stage for classification process. The results show an accuracy of 99.04% and sensitivity of 100% achieved by SVM to overcome the other works presented in the literature.

This research shows that the proposed hybrid system can be employed as a powerful tool to facilitate final decision of clinical diagnosis and can be successfully applied for various medical data classification. Although the results of this research is promising, number of general directions remain open to extend this work. This research can be extended to investigate other real-world problems of different domains. Also, testing the scalability of the proposed hybrid system is an interesting subject. Furthermore, the dataset used in this research is a well-known two-class problem; another future work can be evaluating the performance of the proposed system on other multiclass problems.

References

[1] Russell, S., Norvig, P., & Intelligence, A. (1995). A modern approach. Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs, 25, 27.I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[2] Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in medicine, 23(1), 89-109.

[3] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

[4] Rumelhart, D. E., Hinton, G. E., Williams, R. J. (1986), Learning internal representations by error propagation. In: Rumelhart D E, McClelland J L et al. (eds.) Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge, MA, 1: 318-362.

[5] Liu, H. and H. Motoda (2001), Instance Selection and Constructive Data Mining, Kluwer, Boston.

[6] Kohavi, R. (1996, August). Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In KDD (96), 202-207.

[7] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

[8] Kuncheva, L. I. (2004). Combining pattern classifiers: methods and algorithms. John Wiley & Sons.

[9] Jain A, Dubes R (1988) Algorithms for clustering data. Prentice-Hall, Inc, Upper Saddle River

[10] Bezdek, James C. Pattern recognition with fuzzy objective function algorithms. Springer Science & Business Media, 2013.

[11] Yager, Ronald R., and Dimitar P. Filev. "Approximate clustering via the mountain method." IEEE Transactions on Systems, Man, and Cybernetics 24.8 (1994): 1279-1284.

[12] Gustafson, Donald E., and William C. Kessel. "Scientific Systems, Inc. 186 Alewife Brook Parlcway Cambridge, Massachusetts 02138." (1979).

[13] Dave, Rajesh N., and Kurra Bhaswan. "Adaptive fuzzy c-shells clustering and detection of ellipses." IEEE Transactions on Neural Networks 3.5 (1992): 643-662.

[14] Xu R, Wunsch D (2005) Survey of clustering algorithms. IEEE Trans Neural Netw 16:645–678.

[15] Bashir, S., Qamar, U., & Khan, F. H. (2015). Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble. Quality & Quantity, 49(5), 2061-2076.

[16] Karabatak, M. (2015). A new classifier for breast cancer detection based on Naïve Bayesian. Measurement, 72, 32-36.

[17] Lavanya, D., & Rani, K. U. (2012). Ensemble decision tree classifier for breast cancer data. International Journal of Information Technology Convergence and Services, 2(1), 17.

[18] Salama, G. I., Abdelhalim, M., & Zeid, M. A. E. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer (WDBC), 32(569), 2.

[19] Aruna, S., Rajagopalan, S. P., & Nandakishore, L. V. (2011). Knowledge based analysis of various statistical tools in detecting breast cancer. *Computer Science & Information Technology*, 2, 37-45.

[20] Lavanya, D., & Rani, D. K. U. (2011). Analysis of feature selection with classification: Breast cancer datasets. Indian Journal of Computer Science and Engineering (IJCSE), 2(5), 756-763.

[21] Christobel, A., & Sivaprakasam, Y. (2011). An empirical comparison of data mining classification methods. International Journal of Computer Information Systems, 3(2), 24-28.

[22] Hu, J., Wu, W., Zhu, B., Wang, H., Liu, R., Zhang, X., ... & Tian, C. (2016). Cerebral Glioma Grading Using Bayesian Network with Features Extracted from Multiple Modalities of Magnetic Resonance Imaging. PloS one, 11(4), e0153369.

[23] Spasojević, S., Santos-Victor, J., Ilić, T., Milanović, S., Potkonjak, V., & Rodić, A. (2015, July). A Vision-Based System for Movement Analysis in Medical Applications: The Example of Parkinson Disease. In International Conference on Computer Vision Systems, Springer International Publishing, 424-434.

[24] Agrawal, P., & kumar Dewangan, A. (2015). A Brief Survey on the Techniques used for the Diagnosis of Diabetes-Mellitus. International Research Journal of Engineering and Technology (IRJET), 2(03), 2395-0056.

[25] Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 27.

[26] M. Abdullah, F. AlAnzi, S. Al-sharhan, (in press). Efficient Fuzzy Techniques for Medical Data Clustering, Proc in The 9th IEEE-GCC Conference and Exhibition (GCCCE), Bahrain, 2017, PP. 400