A COMPARATIVE ANALYSIS OF CLASSIFICATION TECHNIQUES ON MEDICAL DATA SETS

Pooja Mittal[1], Nasib Singh Gill[2]

[1]*Department of Computer Science & Application, Maharshi Dayanand University, Rohtak 124001, Haryana, India*
[2]*Department of Computer Science & Application, Maharshi Dayanand University, Rohtak 124001, Haryana, India*

Abstract

Classification is the preliminary stage of data mining which is used to categorize the dataset in smaller groups where each group contains the similar data items. The classification basically deals with two main parameters; one is the number of classes and another is the criteria for deciding the class members. Different recognition algorithms also use the classification process as an initial stage to perk up the efficiency and the accuracy. The accuracy of the classification algorithm also decides the effectiveness of its use in other mining applications. The present work is about to analyze the effectiveness of the most popular classification techniques. In this paper, the analysis has been performed for five different classification algorithms in terms of accuracy, kappa statistics, execution time, mean absolute error under three datasets, collected from medical domain. The work has been implemented in WEKA environment and obtained results show that SVM is the most robust classification method and KNN is the least effective classifier for medical data sets.

Keywords: *KNN, Neural Network, SVM, Decision Tree, Classification.*

----------***----------

1. INTRODUCTION

Data mining on medical data is a challenging area. It is a process of deriving consequential and imperative information hidden in exhaustive comprehensive data. Classification is an integral component of clinical data mining as it synthesizes the bulky data into meaningful groups, and assists in effective mining. Classification [1] is an initial stage of data mining to divide the data in sub groups according to the data and dataset features. These features can be application based or the data based. Variety of Classification algorithms is available like KNN, Decision Trees, Bayesian networks, SVM. Initially, the classification algorithm divides the dataset in two parts called training and testing data. The training data is the input data that is been learned under by the classification algorithm as a premature stage. The learning process can be supervised or unsupervised. Once the learning of training data is accomplished, the classes are identified. The testing data is a new entity for which the class is unrecognized. This testing data is analyzed under different parameters and compared with available class's features. The maximum harmonized class is taken as the data class [2]. Classification algorithms like decision tree, neural network, support vector machine, KNN etc [2,3], collectively forms a generalization model to perform the classification. Some of these methods are statistical and some are soft in terms of rules and works on weighted values. In section 3, these methods are described in detail.

2. LITERATURE SURVEY

An early detection & intervention plays a significant role in controlling the clinical deterioration of ICU patients [11]. An integrated data mining approach was designed to give early warnings and alarms by synthesizing large set of features like DFA, Time Series, Entropy by applying linear and non-linear classification, forward feature selection and exploratory under sampling. An improved hybrid prediction model was proposed by ILango & Ramaraj[12] by implementing F-Score feature selection to deduce optimal feature set from high dimensional data bases. They achieved predictive accuracy of 98.9427% for diabetes data set. Data mining techniques are significantly important to derive and concise information from large & high dimensional clinical data [13]. Real data sets from various domains were analyzed & identified the performance of various techniques, to pick out the best out performer. Quantum C4.5 and Random tree produced 100% accuracy on one data set and 91.36% on another data set. Accuracy of any technique depends on the algorithm and on the nature & behavior of data set. The medical coding problem can be visualized as multi label classification problem, in which patient's information is translated in standard pre-defined codes. A multi label large margin classifier is capable of learning the code structure, using the previous knowledge [14]. Medical data mining primarily focuses on hidden pattern extraction. An amalgam model was developed for classifying medical data set .Combination of multistep pre-processing, k-means and K nearest neighbor improves the performance of the process [15]. Missing values present in medical data set effects the pattern extraction process significantly. Imputation is the most popular and common approach for solving MV's problems. Heart failure data set was referred and concluded that no universal imputation tool is available which can outperform even if applied on diverse data sets [16]. Temporal patterns can be mined by applying fuzzy

neural networks. Lower approximations can be derived by implementing hypothesis and fuzzy decision tables [17]. A novel approach was proposed for classification based on bijective soft sets. Bijective soft set theory performs the analysis on data and identifies the dependency in data values. It also discovers the redundant information and identifies the classes over data values . It comes out to be a valuable asset for inductive learning. Precision, Recall and F-measure were the factors which were analyzed for comparing the performance of bijective classification from decision table and naïve bayes[18]. Neural networks when applied on data set of heart disease patients resulted in sensitivity of 81.1%, specificity of 78.7% and accuracy of 80.2%. Decision tress yielded sensitivity, specificity and accuracy of 81.7%, 76.0% and 79.3% where as logistic regression achieved 81.2% , 73.1% and 71.1% of sensitivity, specificity and accuracy respectively [19]. Various medical data sets from diverse domains were analyzed by applying SSI, KD, RB and L measures. SSI produced maximum classification accuracy on different data sets [20],[16]. Maintenance of patients data sets is equally important as it may assist in acquiring knowledge and identifying problems [21],[12]. Heart disease diagnostic is one of the prime applications of data mining in medical stream. Many researchers contributed to develop intelligent systems. My Chau Tu, Dongil Shin applied decision tree C4.5, bagging with C4.5 and bagging with Naïve Baye's theorem and analyzed the effectiveness, correction rate. 10 fold cross validation was used for evaluation. Out of these techniques bagging with Naïve Baye's theorem out shines with 82.5% accuracy [22]. Pre processing improves the performance of mining techniques by removing noisy data, substituting missing values (MV's) [15][16] whereas k-means algorithms are used to eliminate incorrect classification. Value of k depends on the nature and behavior of data. Larger the value of k, lesser the chances for noise. 97.4% accuracy was achieved by caching KNN & k means. The state of classifiers are mainly decided by the nature of the data set, when k-fold cross validation was applied [23]. Dempster – Shafer's theory of evidence combination was implemented on two varied medical domains of Breast Cancer and Skin lesions, by applying k-nearest neighbor, Naïve Bayesian and Decision Tree. In this scenario, beliefs were computed by classifier outputs. Belief m(X) is a measure of trust or confidence with $\sum_{X \in \Omega} m(X) = 1$, where $m(\phi) = 0$ where ϕ is empty [23]. Classification can be improved by using evidence combination approach. To enhance the performance, the probability theory is modified and the mathematical theory of Evidence is devised for handling uncertainty [24]. Binary – coded genetic algorithms and Real-coded genetic algorithms are used for assigning weights to the features, so that set of optimal features can be deduced from high dimensional data. Different k-NN algorithms (crisp k-NN, fuzzy KNN and weighting fuzzy k-NN) are evaluated and compared on same data set. They achieved 98.77% of accuracy by applying RGA based model whereas GA approach comes out to be time consuming [25]. Knowledge can be extracted by clustering the given data set in to soft clusters and can be fused by using serial and parallel fusion to outperform as classifier [26]. Hybrid approaches can be used for classifying the given voluminous data [27,26,18]. Ant Colony Optimization algorithm is an efficient approach used for classification purpose[27]. Quality can be improved if ACO is combined with mRAR, a feature selection algorithm.

3. CLASSIFICATION APPROACHES

Classification is a form of data analysis [10], which is used for extracting a model for describing and differentiating the data classes of given data objects, with an objective of predicting the class for an object whose class label is not known. The classification process can be broadly divided into two phases: Learning step (training data) and Classification step(testing data). Classification is used to predict categorical labels including discrete and unordered values. This derived model can be represented in many forms known as classification algorithms such as IF-THEN rules, decision tree, mathematical formulae, neural networks etc. Some of the popular techniques are described below.

3.1 Decision Tree

It is the hierarchical decision making approach used to partition the dataset. It is an approximate discrete function technique for retrieving useful expressions.

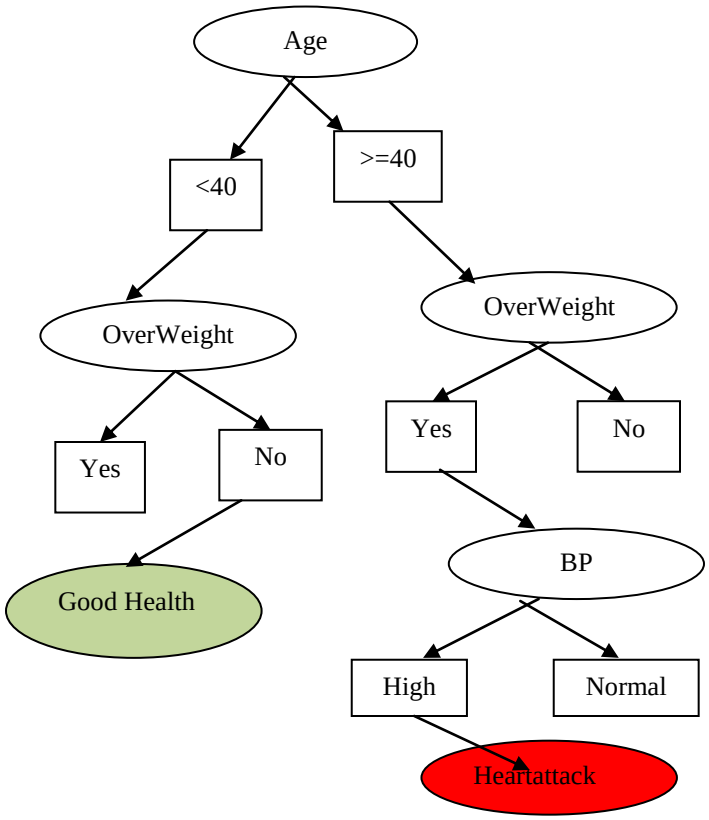

Fig. 1 A simple classification or decision tree

According to this approach, the dataset is classified to n mutual exclusive datasets and each dataset will be defined with a label. Now the data points are identified by a certain action to its relative data class. Decision Tree is a supervised learning technique that defines a transparent tree based structure to perform the action on available dataset. The tree structure itself contains number of nodes that are connected by the edges. These nodes basically define the conditions and the edges define the relative event based on the condition. Each edge itself defines a separate class. The decision is here been taken respective to the true or false case of a conditional analysis. If the condition is true, one category will be elected otherwise second categorization. Figure 1 is showing the simple classification process defined by decision tree [3,4].

3.2 Bayesian Classification

Bayesian network, the probabilistic graphical model that identify the relation between the variables and based on the dependency analysis, the classification will be performed. The Bayesian network is a directed acyclic graph based model. The dependency is here been identified between different attributes in terms of conditional analysis. Each attribute is at first defined independently and with the next level, the relation between these attributes is identified. Each attribute itself is defined with some weight age called the probabilistic analysis and as the relations are identified, the conditional probability is evaluated for each factor. This classifier performs the learning on training data under the conditional probability analysis on each attribute as well assign a random class label to each instance. Once the conditional probability based acyclic graph is generated, the next work is to predict the class with highest probability vector. The goal of classification process is to predict the discrete value of class for any testing data. The structure of Bayesian network is been defined under the one to one feature vector. The arcs defined in the graph represents the features based on which the conditional dependencies are evaluated [6].

3.3 Neural Network

Neural network is one of effective soft computing based classification algorithm that uses the concept of neurons that logically represents the working of human brain. In this classification process, the data values are represented by the neurons and the connectivity is represented by synapses. It is basically the layered approach in which there are two main layers called two end points represented by input and output layer. Other then these two layers, the model also have m intermediate layers called hidden layers. On each layer some weight age is assigned. The graphical representation of Neural network is given in figure 2. As we can see, the middle layer of the network defines the weights to different input values so that effective classification will be done. Neural network accepts the dataset as input layer and represent it as the network nodes. The predictor weights are applied to these nodes in hidden layer. This layer actually defines the degree of connectivity between the nodes. After adjusting the weightage, output layer is derived as the final result [6,7].

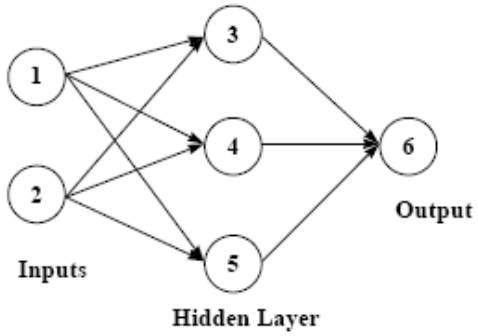

Fig. 2: A Three Layer Neural Network

3.4 Support Vector Machine

SVM is another robust and successful classification algorithm. SVM basically works as the linear separator between two data points to identify two different classes in the multidimensional environment. The prime objective of this approach is to maximize the margin between the classes and to minimize the distance between the hyper plane points. SVM basically defines the dealing of interaction respective to the features and the repetitive features. SVM split the dataset in two vector sets under n dimensional space vector. The SVM algorithm basically construct a hyper plane environment so that each element is been compared respective to the separated linear line. Hyper-plane concept is presented to perform the data separation based on largest distance analysis to identify the classes. To reduce the error ratio, the largest margin classifier is defined. The work also includes the analysis based on margin vector along with support vector analysis [8].

3.5 k-Nearest Neighbor

KNN is the instance based statistical analysis approach to perform data classification, called as lazy learning algorithm. It is the simplest algorithmic approach among all algorithmic approaches. According to this approach, an object is classified by neighbor point analysis based on majority analysis. The object that will get the highest vote will be selected as the class member respective to the distance defined class. Once the classification rule is decided, the relative neighboring objects are identified. If the value of k=1, then it is simply called as nearest neighbor. K-NN requires

1. An integer k
2. A training data set
3. A metric to measure closeness

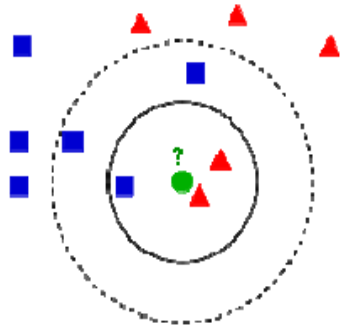

Fig. 3 Example of KNN classification

Example of KNN classification is shown in figure 3. Here input data objects are represented by greed circles and the object classes are identified by blue and red objects. In case of three classes, triangles are represented by one class and rest objects are represented by other objects. In case of 5 classes, three classes are presented by square and rest two classes are represented by triangles shown inside the outer circle[9,10].

4. EXPERIMENTAL RESULTS AND ANALYSIS

The performance analysis was carried on five different algorithms for three different datasets. The datasets considered here are from medical domain. The classification algorithms used in this work are described in previous section. The present work has been implemented in WEKA (Waikato Environment for Knowledge Analysis) environment [4,9] and the results have been taken under different parameters: the accuracy, Kappa Statistics[4], Mean Absolute Error and Execution Time. The results obtained from these different models have been defined in the form of tables as well as graph.

4.1 Datasets

In this work, three medical datasets are considered, which are collected from the UCI Repository. These datasets are effective enough to show the classification process. These datasets are here analyzed under different classification parameters. These datasets are available in ARFF format. The detailed description of these datasets, in terms of features and data points, is given in table 5.

Table 5: Medical Datasets

Sr. No.	Dataset Name	Instances	Attributes
1	heart-statlog.arff	270	14
2	diabetes.arff	768	9
3	hepatitis.arff	155	20

The medical datasets that have been utilized to conduct the classification are taken from diverse range of medical areas, to ensure the excellence of analysis . Every dataset has different types of data, including numbers, text and other domain data points. Each of the dataset is explored explicitly due to their uniqueness in terms of their varying attributes, discrete or continuous nature of data etc. These datasets are been analyzed for the classification task by using WEKA tool under different classification approaches. To perform this classification, 10-fold method is adopted in this present work. WEKA is an open source data mining software tool written in java. WEKA, itself contains number of built-in data mining algorithms so that different mining operations can be performed directly. WEKA is used by the researchers to analyze the effectiveness of different machine learning algorithms. In this present work, we have used WEKA to perform the analytical study of classification algorithms on medical datasets. Various parameters considered, are described next.

4.2 Accuracy Analysis

Accuracy of a classification algorithm is been defined in terms of number of correctly classified instances. Accuracy Analysis is given by

$$\text{Accuracy Analysis} = \frac{\text{CCount}}{\text{TCount}} \times 100$$

Here,
CCount is Number of Correctly Identified Objects, and
TCount is Total Number of Objects.

Higher the accuracy level, more effective the algorithm will be.

Table 1: Accuracy Analysis of Different Classification Algorithms

Dataset name	**Classification Techniques**				
	Bayesian Networks	Neural Networks	SVM	KNN	Decision Tree
heart-statlog.arff	83.7037	77.4074	84.0741	75.1852	76.6667
diabetes.arff	76.3021	75.1302	77.474	69.7917	74.2188
hepatitis.arff	83.2258	80	85.1613	80.6452	81.2903

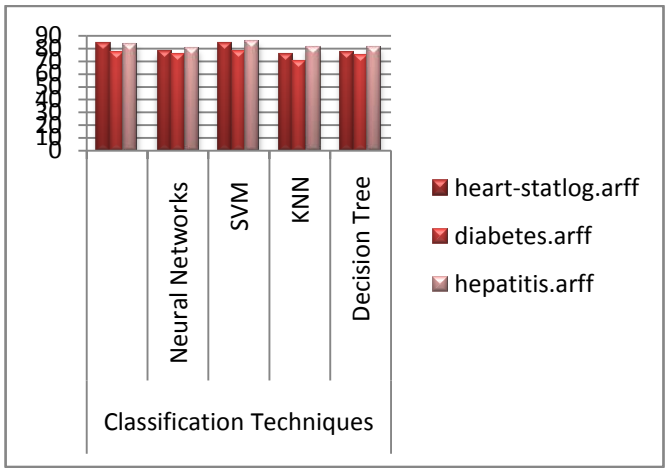

Fig 4: Accuracy Analysis of Different Classification Algorithms

As it can be viewed, figure 4 is showing the accuracy based comparison of different algorithms. The figure shows that the SVM is most robust, effective and consistent classifier for different datasets. SVM provided the highest accuracy among all algorithms whereas KNN is the least effective algorithm in terms of accuracy analysis.

4.3 Based on Execution Time

To analyze any algorithmic approach, the execution time is one of the foremost parameter. In this present work, we have analyzed the execution time to identify the efficient classification algorithm. Here table 2 is showing the execution time results obtained from different algorithms.

Table 2: Execution Time Analysis of Different Classification Algorithms

Dataset name	Classification Techniques				
	Bayesian Networks	Neural Networks	SVM	KNN	Decision Tree
heart-statlog.arff	0.02	1.3	0.08	0.003	0.03
diabetes.arff	0.02	2.03	0.11	0.005	0.06
hepatitis.arff	0.04	1.28	0.09	0.004	0.03

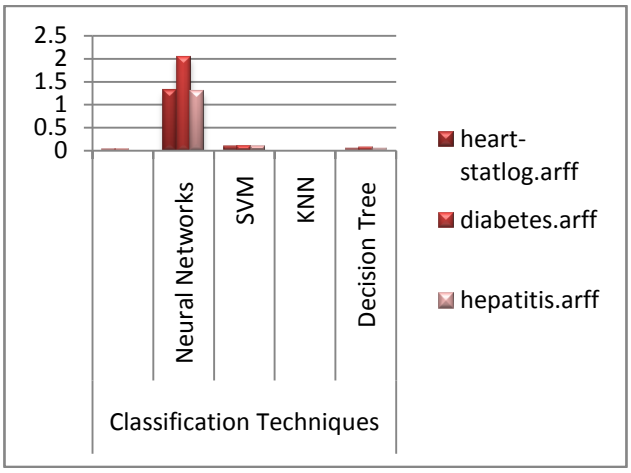

Fig 5: Execution Time Analysis of Different Algorithms

Here figure 5, is showing the execution time analysis of different classification algorithms. As we can see, the KNN is most efficient classification algorithm whereas the results obtained for neural network are worst.

4.4 Based on Kappa Statistic

Kappa Statistics is the statistical analysis based on the inter-rater agreement for qualitative data. It basically performs the analysis between different classes. Higher Value of kappa statistic is considered as good. Here figure 3 is showing the comparative analysis of different algorithms under the kappa statistics.

Table 3: Kappa Statistic Analysis of Different Algorithms

Dataset name	Classification Techniques				
	Bayesian Networks	Neural Networks	SVM	KNN	Decision Tree
heart-statlog.arff	0.6683	0.5444	0.6762	0.4988	0.5271
diabetes.arff	0.4664	0.4445	0.4708	0.3223	0.4246
hepatitis.arff	0.5107	0.3825	0.5309	0.3953	0.394

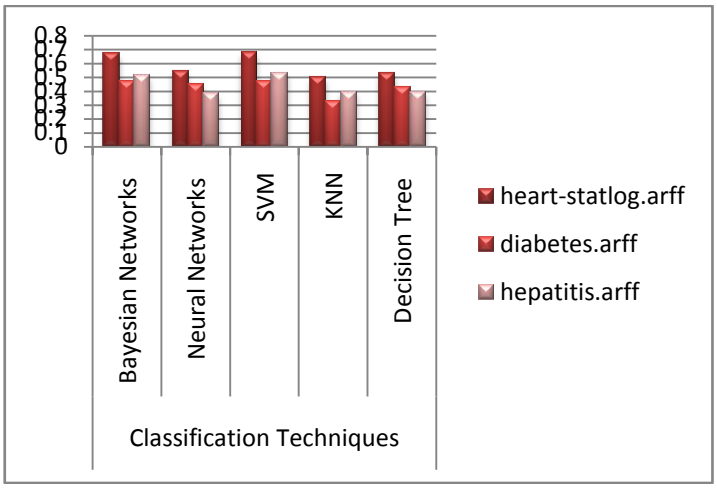

Fig 6: Kappa Statistic based Analysis of Different Classification Algorithms

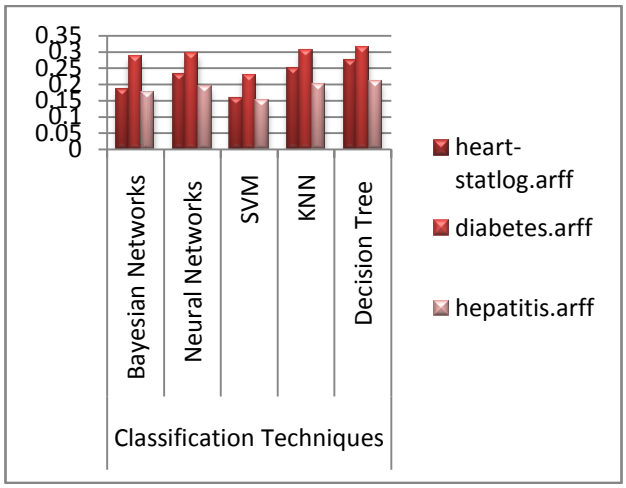

Fig 7: Mean Absolute Error based Analysis for Different Algorithms

As we can see, figure 6 shows the Kappa statistics based analysis. As we can see, SVM classification method has provided the highest Kappa statistic value for different dataset. It shows the effectiveness in terms of robustness. The values obtained in case of KNN algorithm is least that shows the least effective algorithm under this parameter.

4.5 Based on Mean Absolute Error (MAE)

MAE actually forecasts the capability of an algorithm. Lesser the MAE , higher the capability of the algorithm to perform the prediction. Here table 4, is showing the results obtained from different algorithm s under mean absolute error parameter.

Table 4: Mean Absolute Error based Analysis for Different Classification Algorithms

Dataset name	Classification Techniques				
	Bayesian Networks	Neural Networks	SVM	KNN	Decision Tree
heart-statlog.arff	0.1835	0.2328	0.1593	0.2502	0.274
diabetes.arff	0.2841	0.294	0.2253	0.3027	0.3134
hepatitis.arff	0.1754	0.1928	0.1484	0.1979	0.2073

Here figure 7, is showing the estimation of different algorithms under the mean absolute error. As we can see, MAE value in case of SVM algorithm is least that shows the accurate prediction capability of the algorithm. Whereas the highest values driven from the KNN algorithm, shows that the algorithm is not effective for the prediction.

5. CONCLUSIONS

In this paper, the analysis of different classification algorithms has been performed under four different parameters called execution time, mean absolute error, kappa statistic and accuracy analysis for three medical datasets. The obtained results show that the SVM is the most robust, consistent and reliable classification algorithm whereas KNN is the worst algorithm for the classification.

REFERENCES

[1] Gupta, M., and Aggarwal, N. , "Performance Analysis of Classification Techniques on XML Dataset", International Journal of Computer Science and Technology IJCST Vol. 1, Issue 1, pp. 76-79, 2010.

[2] Justin, T., Gajsek, R., Struc, V., and Dobrisek, S., "Comparison of Different Classification Methods for Emotion Recognition", MIPRO 2010, Opatija, Croatia, pp. 700-703, 2010.

[3] Gupta, S., Kumar, D., and Sharma, A., "Data Mining Classification Techniques applied for Breast Cancer Diagnosis and Prognosis", Indian Journal of Computer Science and Engineering (IJCSE) Vol. 2 No. 2, pp. 188-195, 2011

[4] Viera, A. J., and Garrett, J. M., "Understanding Interobserver Agreement: The Kappa Statistic", Research Series Vol. 37, No. 5 , pp. 360-363, 2005.

[5] Desai, A., and Rai, S., "Analysis of Machine Learning Algorithms using WEKA", International Conference &

Workshop on Recent Trends in Technology, (TCET) 2012 Proceedings published in International Journal of Computer Applications (IJCA) 27, pp.27-32, 2012.

[6] Kumari, Milan, and Godara, Sunila, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", International Journal of Computer Science and Technology Vol. 2, Issue 2, pp. 304-308, 2011.

[7] Ture, M., Kurt, I., Kurum, A. T., and Ozdamar, K., **"**Comparing classification techniques for predicting essential hypertension**",** Expert Systems with Applications 29, pp. 583–588, 2011.

[8] Burges, C., "A Tutorial on Support Vector Machines for Pattern Recognition". Data Mining and Knowledge Discovery, Vol. 2, pp. 121-167, 1998.

[9] Pushpa, "Comparision of Clustering Techniques using WEKA", M. Tech. Thesis, Guru Jambheshwar University of Science and Technology, Hisar, India,2010.

[10] Han, J., and Kamber, M., "Data Mining: Concepts and Techniques", 2nd edition, Morgan Kaufmann Publishers, 2006

[11] Yi Mao, Wenlin Chen, Yixin Chen, Lu Chenyang, Kollef Marin & Thomas C., " An Intergrated Data Mining Approach to Real Time Clinical Monitoring and Deterioration Warning",KDD'12 ACM 978-1-4503-1462-6/12/08.

[12] Ilango B.S, Ramaraj N., " A Hybrid Prediction Model with F-score Feature selection for TypeII Diabetes Databases", A2CWiC 2010, September 16-17, 2010 .

[13] Jacob S.G. , Ramani R.G, " Mining of Classification Patterns in Clinical Data through Data Mining Algorithms", ICACCI'12 –ACM 978-1-4503-1196-0/12/08.

[14] Yan Yan, Fung G., DY J.G, " Medical Coding Classification by Leveraging Inter-Code Relationships", KDD'10 July 25-28 2010 Washington, USA.

[15] NirmalaDevi M, Balamurugan S, Swathi U V, " An amalgam KNN to predict Diabetes Mellitus", 2013 IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology.

[16] Davis D N, Zhang Y, Kambhampati C, Goode K, Cleland J.G.F, " A Comparative study of missing value imputation with multi class classification for clinical heart failure data", 2012 IEEE, 9th International Conference on Fuzzy Systems and Knowledge Discovery.

[17] Sethukkarasi R., Keerthika U., Kannan A. , " A Self Learning Rough Fuzzy Neural Network Classifier for Mining Temporal Patterns", ICACCI'12 –ACM 978-1-4503-1196-0/12/08.

[18] Kumar S U, Inbarani H, Senthil Kumar, " Bijective Soft Set Based Classification of Medical Data", 2013 IEEE, International Conference on Pattern Recognition, Informatics and Mobile Engineering.

[19] Khemphila A, Boonjing V, " Comparing performances of logistic regression, decision trees and neural networks for classifying heart disease patients", 2010 IEEE , International Conference on Computer Information Systems and Industrial Management Applications.

[20] Saastamoinen K, Ketola J, " Medical Data Classification using Logical Similarity based Measures", 1-4244-0023-6/06 2006 IEEE.

[21] Mutalib S., Razak Abd. R., Nordin S., Rahman S.A., Mohamed A., " Intelligent classification in Medical Data",2012 IEEE International Conference on Biomedical Engineering and Sciences.

[22] Tu M C, Shin D, Shin Dong, " A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms", 2009 8th IEEE International Conference on Dependable, Autonomic and Secure Computing.

[23] Aslandogan Y.A, Mahajani G. A, " Evidence Combination in Data Mining", 2004 IEEE Proceedings of the International Conference on Information Technology: Coding and Computing .

[24] Shafer G, " A mathematical theory of evidence", Princeton University Press, 1976.

[25] Tang P.H, Tseng M.H, " Medical data mining using BGA and RGA for weighting of features in Fuzzy k-NN classification", 2009 8th IEEE International Conference on Machine Learning and Cybernetics.

[26] Hassan S. Z, Verma B, " A Hybrid Data Mining Approach for Knowledge Extraction and Classification in Medical Databases", 2007 IEEE 7th International Conference on Intelligent Systems Design and Applications.

[27] Michelakos I, Papageorgiou E, Vasilakopoulos M, " A hybrid classification algorithm evaluated on medical data", 2010 IEEE Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises.