Empirical Performance Analysis of Decision Tree and Support Vector Machine based Classifiers on Biological Databases

Muhammad Amjad[1*], Abid Rafiq[3]
Department of Computer Science and Information Technology
University of SargodhaSargodha, Pakistan

Zulfiqar Ali[2]
Department of Computer Science and Information Technology
University of Central Pujab (UCP), Lahore, Pakistan

Nadeem Akhtar[4], Ali Abbas[6]
Department of Computer Science and Information Technology
The University of Lahore (UOL), Lahore, Pakistan

Israr-Ur-Rehman[5]
Department of Computer Science
Islamia College University, Peshawar, Pakistan

***Abstract*—The classification and prediction of medical diseases is a cutting edge research problem in the medical field. The experts of machine learning are continuously proposing new classification methods for the prediction of diseases. The discovery of classification rules from medical databases for classification and prediction of diseases is a challenging and non-trivial task. It is very significant to investigate the more promising and efficient classification approaches for the discovery of classification rules from the medical databases. This paper focuses on the problem of selection of more efficient, promising and suitable classifier for the prediction of specific diseases by performing empirical studies on bunch mark medical databases. The research work under the focus concentrates on the benchmark medical data sets i.e. arrhythmia, breast-cancer, diabetes, hepatitis, mammography, lymph, liver-disorders, sick, cardiotocography, heart-statlog, breast-w, and lung-cancer. The medical data sets are obtained from the open-source UCI machine learning repository. The research work will be investigating the performance of Decision Tree (i.e. AdaBoost.NC, C45-C, CART, and ID3-C) and Support Vector Machines. For experimentation, Knowledge Extraction based on Evolutionary Learning (KEEL), a data mining tool will be used. This research work provides the empirical performance analysis of decision tree-based classifiers and SVM on a specific dataset. Moreover, this article provides a comparative performance analysis of classification approaches in terms of statistics.**

Keywords—Classification; rules discovery; support vector machine; decision tree

I. Introduction

The Knowledge Discovery is processing of finding the non-trivial, useful and hidden patterns from a very large database. Knowledge discovery and data mining are a new trend in information technology. Traditionally a large part of the process was done by manually that is time-consuming task. With time new technologies invented and task shifted from manually to computerized form. Business knowledge is necessary in advance to compete in the world. Data storage is now a day reached to amount of terabyte size [1]. But it is necessary to extract useful knowledge from it for use. So knowledge discovery is the name of the discovery of hidden knowledge from large databases. Knowledge discovery contains the steps of data preparation, data preprocessing, and hypothesis generation, the formation of the pattern, evaluation, knowledge representation, knowledge refinement, and knowledge management [2]. It also includes many stages for databases updating.

Machine Learning methods and biological databases play a significant role in disease diagnosis. It helps in future for diagnosing of medicine. The biological database includes information about gene structure, function, and similarities of structure and sequences of biological data. Classification of the biological database can be done in two forms as a specialized and comprehensive database. The comprehensive database includes different species database, for example, GenBank [3] and specialized databases consist of a special organism or species databases, for example, WormBase [4].

Machine learning becomes a necessary part of solving the problem in every branch of science. In biomedicine to predict genetic sequence and protein structure machine learning has been used [5]. Machine learning is used to extract hidden knowledge for the different data set. It includes neural network, boosting, support vector machine and decision trees [6]. In machine learning, two ways are performed for data mining. It is supervised learning we make a dataset to extract new data from a large amount of data. New data and training data set match for validation of result. But in unsupervised learning, some pattern is used to classifying the data without explicit instruction [7]. Reinforcement learning focus on the reward and output achieve in the form of reward and punishment. An agent is required to gain the maximum reward to gain the result. Agent focuses on the positive situation to gain maximum reward. Negative situation decreases the reward. This type of learning is used in control theory, statistics, information theory, etc.

This research article investigates the performance of Decision Tree approach and Support Vector Machine Algorithm for the discovery of classification rules. The

*Corresponding Authors.

interesting and useful discovered rules are used for the building of classifiers. The classifiers are applied for the diagnoses of the various harmful diseases. In this paper, we use KEEL [8] data mining tool for the data processing and classification of the biological databases.

Section II provides the related work published in contemporary literature. Section III gives information about the decision tree-based classification and provides the empirical performance analysis of selected classifiers on medical databases. Section IV provides a basic understanding of SVMs and comparative empirical study on medical data sets. Sections V and VI provide the experimental setup and discussion on the results produced during the under focused research study and the last section concludes the findings of the research work.

II. Related Work

This section provides the literature review of the various research carried by the different researchers in this field. The following section gives information about the use of different classification for the discovery of rules and the classification of different biological diseases.

There are many techniques are used to find a pattern inpatient health data. The best system is one that is the efficient, adoptive, generic and affordable system. Many factors affect the result of analysis like an error in online databases, sensor's settlement. This study shows that ASP logic approach is the best use for incomplete biological data. Artificial Neural Network is best used for single purpose system. ANN generates best better result than ASP and another approach used in the health care system. If the hardware is costly then it difficult to use this system [9].

There are many data mining algorithm available but this study provides a comparative study of three algorithms Naïve Bayes, Decision Tree and Multi-Layer Perceptron Neural Network. In this study, window operating system 8.1 is used with WEKA data mining tool. Ebola Disease data set contain the range of 250-10000 instances that are stored in MySQL. According to this study, the Naïve Bayes algorithm shows a negative correlation, with the increase in the dataset it performance lead to a decrease. WEKA shows a positive correlation. Naïve Bayes is the best and popular machine learning algorithm is fast in training [10].

Mohammed H. Tafish and Dr. Alaa M. El-Halees proposed a model as Breast Cancer Severity Degree Predication Using Data Mining Techniques in the Gaza Strip described that in Gaza Area cancer disease and diabetes growth are top disease during the last decades. They used a data mining method to diagnose cancer and diabetes disease. They proposed a model using data mining techniques like SVM, KNN, and ANN. Breast cancer data taken from Gaza hospital used, after evaluation and test by applying the above techniques they obtain 77% accuracy for the prediction of the severity of breast cancer [11].

Manickam Ramasamy at el. proposed a model for predicting hepatitis in which they provide an empirical analysis of the decision tree algorithm by using Hepatitis data set taken from the UCI machine learning repository. They used different classification algorithm and accuracies of classification are performed by 10 cross validation techniques. By using different classifier they concluded that Random forest takes less running time with the highest accuracy of 87.50%. This accuracy gives help in ailment prediction and classification in the field of medical science [12].

In this study extended deep learning method is used for classifying multimedia data set. Convolution Neural Network is a deep learning method is costly but this paper feed low level features in this approach. To find the best result CNN is used with the bootstrapping method. TRECVID data set is used in this approach which is high-level imbalanced data set. This approach works effectively on the use of low-level features that reduced the training time in deep learning [13].

Anuj et al. describe Parkinson's disease. It is the connection between speech impairment and Parkinson's disease. In this paper classification based deep learning (Deep Neural Network, Dimensionality reduction techniques) and machine learning algorithms (Logistic regression, Naïve Bayes, K-Nearest Neighbor, Decision tree, Random forest) are employed with the use of Dimensionality reduction. The data set Parkinson's Speech is used in this approach that is obtained from the UCI machine learning repository. The result is extracted with the base of accuracy. KNN produced 95% highest accuracy with 10 features [14].

Sara Belarouci et al. propose meta-heuristics optimization methods for improvement of medical classifier performance. They are used many algorithms like Genetic Algorithm PSO, Simulated Anneeling to compare with Least Square Support Vector Machine to improve the classification with aspect to False Positive and Negative. Meta-heuristics Optimization is best for solving the problem of unbalance dataset. Five different datasets related to various diseases like Liver Disorder, Appendicitis, and Diabetes. This approach will help doctors to diagnose many diseases effectively [15].

Tharaha S and Rashika K proposed this research using Hybrid Artificial Neural Network and Decision Tree algorithm for disease recognition. They used Artificial Neural Network for training data and decision tree for classification of data because the Decision Tree algorithm is a good classifier. Datasets are taken from the human blood detecting and sensor counting, stored with different attributes. Time taken for test split in ANN is 0.09s and where decision tree took time is 0.14s. The result is shown by apply WEKA 3.8.1 version. The combination of these two algorithms gives the best result than separate used and provide the best help for disease diagnosing [16].

Dania Abed aljawad et al. proposed an empirical study of Bayesian Network and Support Vector Machines for Breast Cancer surgery Survivability Prediction. They used Haberman's survival dataset and evaluate the performance of the Bayesian network and Support Vector Machine using WEKA tool. Empirical research shows that Support Vector Machine best performs with an accuracy of 74.44% than Bayesian network with an accuracy of 67.56%, Imbalance data is converted into balance. This study helps the doctors to the prediction of the patient stage of cancer using old data as a sample to new data [17].

P. Hamsagayathri and P. Sampath proposed a Priority Based decision Tree Classifier for Breast cancer. Women mostly from 40-70 age affected with breast cancer. So they proposed a model for prediction of breast cancer. Classification provides a vital role in the detection of breast cancer and helps the researcher to analyze and classify data. SEER breast cancer data set is used in this paper. Two decision tree algorithm J48 and priority-based decision tree algorithm are used. The priority-based algorithm provides the best result with less time consuming to build the model. J48 used repetitive but priority base algorithm not used repletion step and 98.51 accuracies [18].

With the reference of above literature review, the specific medical data sets i.e. arrhythmia, breast-cancer, diabetes, hepatitis, mammography, lymph, liver-disorders, sick, cardiotocography, heart-statlog, breast-w and lung-cancer are not used to investigate the performance of Decision Tree (i.e. AdaBoost.NC, C45-C, CART, ID3-C) and Support Vector Machines. In this research study will Decision Tree based classifiers and SVM Machines for the discovery of classification rules. The problem statement and objectives of this research are given in the next sections.

III. Decision Tree based Classification

After Decision Tree is most popular supervised machine learning algorithm applied for the various classification problems. It is used for classification and regression problems. Decision tree provides the result which is easily understandable by humankind. A decision Tree provide output in a tree-like graph in which each node represents to attribute, each branch provide a rule and each leaf node provide a target class. Target class may be in discrete or in continuous form. Decision Rule may be in IF-then-Else rule. Big decision tree means the more complex rule.

Decision Tree is used as a top-down approach for making a decision tree. It begins from the root node to the leaf node. The decision is made on each internal node where attributes are split into further node if it contains information that can be divided further. More information leads to further classification. If a node cannot have information more then it considered as leaf node that refers to the target value.

Different methods are used to construct a decision tree. Every method used different information for the construction of a decision tree. Large decision tree not considered an accurate and efficient decision tree. Different research shows that the best decision tree is as small as possible. It based on the proper selection of attributes. Attributes selection measures are used to split attributes into further sub attribute. It is a recursive approach. Attributes selection measure checks the impurity of the attribute. Impurity measurement method includes Gain Ratio, distance measures, Gini-index and information gain. ID3, C4.5 focused information gain and CART use Gini-index for attributes selection.

A decision tree process can be divided into two steps: one constructs a decision tree and other to pruning a decision tree. Data mining works on real world data. Data may have some missing value, wrong value, containing noise or even less essential data, so this problem may lead to over-fitting and will destruct the predictive performance. There are two basic strategies for pruning the decision tree i.e. first forward pruning means pruning before completion of decision tree and other post-pruning means pruning after making a decision tree. So forward pruning stop the pruning process before reaching its maturity level and in a post-pruning button-up, approach is used to cut off the node. The Minimum Description Length Principle, Expected Error Rate Minimization Principle and Principle of Occam's Razor are used for pruning.

A. ID3

ID3 stands for Iterative Dichotomize 3. It is built by J.R Quinlan in [19]. It is the core algorithm to build a decision tree. It generates all possible decision tree. It simply classifies the training and testing set for the dataset. It does not require much more computation as compared to another approach for creating a decision tree. It is an iterative approach. It chooses the training set randomly and makes the decision tree. If it answers all object then it terminates the process it not then it add to again in training data for further process. It iterates the process and makes the decision tree correctly up to thirty thousand instance and fifty attributes. This algorithm based on the information gain of candidates attributes. If any attribute has more gain information then it selected for decision tree and less gain information is discorded.

The effectiveness of this approach also depends on the computational requirement based on the gain of untested attributes and non-leaf nodes of the decision tree. The total computational power of the ID3 is relative to the size of the training set, several attribute, and non-leaf nodes. The similarity in attributes extends the computational requirement. In ID 3 time and space are not grow exponentially so it can be used for larger and complex tasks.

ID3 algorithm has some advantages like i.e. easily understandable rule for classification, it is fastest and provides a short tree. It calculation time is a linear function not exponential as well as it has some disadvantages i.e. data may be overfitted or over-classified due to the small sample and for the continuous value it computation time may be more due to make many trees to find where to break the continuum.

B. C4.5

Quinlan et al. proposed the extended version of ID3 that is known as C4.5 in [20]. It is also developed for making a tree. It is developed by Quinlan in 1993. Quinlan described many issues for decision tree-like handling missing value, pruning and converting trees to rule and how C4.5 handle it. Decision tree algorithms used some cases and make a tree-like structure in which the main node is called the root node and other node are test node and leave node. Every decision node used a test and leave node show the class label.

C4.5 algorithm creates a small, accurate and fast decision tree and it is known as a reliable classifier. These are the best and popular properties for making the classification. This algorithm extracts the best information from a set of cases and takes only one attributes for the test. For this purpose information gain and gain, the ratio is used for the selection of best attributes. Some dataset may contain unknown information so Quinlan used C4.5 approach. Information gain for unknown value can be ignored. And known value attribute information gain can be calculated. So information on this test case may be quite small. The unknown value may affect the decision tree making process.

An every decision tree cannot be considered as a good classifier for every data set in respect of making a smaller tree that may not fit for all training data. So avoid by overfitting, many decision tree algorithm used the pruning method. In this method, growing the decision is stopped while deleting the portions of the tree. C4.5 pruning method based on error rate. The error rate of every subtree is calculated if the error rate is low then it will be treated as a leaf node. This process used bottom-up approach. If C4.5 algorithm indicates that tree will be treated as accurate even children of concern node deleted than algorithm considered concern node as a leaf node. If this method proved as good then this decision tree is considered the best decision tree.

Quinlan discusses some shortcoming of c4.5. It has a built-in bias, t take only a single attribute for testing that takes more time computation. It makes the value of the given attribute in the same group and considered as a single value. It may use for single training set once and not used for other training set for binary classification. Suppose attributes for a chemical element that can be classified into the light and heavy element and other training set having an electric conductor that can be classified in conductor and non-conductor. So these groups may overlap with each other. This algorithm cannot is used for both groups. C4.5 used greedy approach for the grouping, so it gives the unsatisfactory result and remains an open problem.

C. Adaboost.NC

AdaBoost.NC is a negative correlation learning algorithm proposed by Wang et al. in [21]. It is used for classification ensemble. AdaBoost.NC algorithm is used for multiclass imbalance data. It provides the solution of two class imbalance problem. AdaBoost.NC provides the best accuracy with random oversampling on the minority class as compared to another balancing approach. The accuracy is achieved by the less border classification and overfitting in the minority class.

AdaBoost.NC is the advance version of AdaBoost for negative correlation but it based on AdaBoost training framework. It provides better classification boundaries and creates lower error correlation as compared to AdaBoost. This is used to improve the performance of the original AdaBoost algorithm. This algorithm is used for better classification in control of upper bound on the generalization error of Traditional AdaBoost. AdaBoost.NC provided the best performance in respect of the distribution of better margin.

AdaBoost is a very simple and effective ensemble algorithm. It is not only used to emphasize to misclassified example, but also provide the mechanism to control the error of misclassification of the same example. Due to this reason, it provides the best accuracy and diversity.

AdaBoost.NC does not show good performance in overall and in minority class working with class decomposition scheme. This algorithm receives and learns from all data information of all classes. It learns from several decomposition problems for partial knowledge. It provides the best performance in analyzing subproblem as compared to combine the whole problem. So it needs to better technique to combine the subproblem to acquire knowledge from AdaBoost.NC.

D. CART

CART stands for classification and regression tree. CART is proposed by Breiman et al. in [22]. It is an algorithm used to construct a decision tree from the categorical and continuous form of data. Classification is used for a categorical form of data and regression tree is constructed from a continuous form of data. The first time Morgan and Sonquist proposed a method to construct a tree by quantitative variable. They gave the name Automatic Interaction Detection. Each cluster is grouped into two clusters. Each predictor is tested on every cluster. Their model naturally incorporates interaction among all predictor.

A classification tree is dependent on discrete or categorical value. Kass (1980) proposed a modification in AID model called CHAID for the creation of a tree from the dependent and independent variable. This model limited to categorical predictor so it cannot be used for the quantitative variable.

These two models have a problem where to stop the tree. Breiman et al. (1984) method show that node that cannot contribute to prediction eliminate from the tree.

CART is a mechanism to construct a decision tree. It makes the solution in a tree-like structure. It starts from the root node and split into a test node on the base of selected attributes. This process ends on the leaf node that cannot be further divided. To make the best and effective tree it used pruning method i.e. Complexity based pruning. Pruning is started from the bottom toward the root node.

CART algorithm may a structure of question and answer of these question lead to the next question. So, the result of these question make a tree structure where to question is not more. CART uses the basic rule for making a decision tree i.e. splitting data rule and stopping rule where the terminal cannot be split and prediction of the leaf node. CART has some advantages like can handle missing value automatically.

IV. Support Vector Machine based Classification

Support Vector Machine was introduced after in the 1990s and used for many engineering application [23]. Support Vector Machine is an algorithm developed for binary classification by Cortes & Vapnik. The objective of this algorithm to find hyper-plane and classification of data points. It is used for separating the two classes with a maximum margin between two points called support vector. SVM algorithm is used for class separation, nonlinearity and overlapping classes where a data point lies in the opponent class [24].

Support Vector Machine classifies the data by using hyper-plane. The hyper-plane can be chosen by either of the sides but optimal hyper-plane is that maximizes the margin between two support vectors. Support vector is the data point that closer to the hyper-plane. Hyper-plane has different features on different location and deleting the support vector can influence the position of the hyper-plane [25].

The main purpose of the Support Vector Machine is to choose the best hyper-plane that classifies the data point correctively with maximum margin. It is easy to find the best hyper-plane in linear form but non-linear hyper-plane is hard to find as compared to linear form. For this purposes, a function called Kernel is used that find the best hyper-plane in no linear form. In non-linear form classification kernel trick, it mapped the input from low dimensional feature space to high dimension feature space.

Support Vector Machine algorithm provides a solution for a limited number of training data in more time and they consume more time for large databases [23]. It is used for text and hypertext categorization, classification of images, image segmentation, and hand-written recognition of character.

The following subsection describes the SVM based classification methods selected for the empirical study in this thesis. The naming convention for the methods is used of KEEL implementations.

A. C-SVM

The C-SVM is a new type of support vector machine proposed by Cortes and Vapnik in [26]. It used non-linear mapping to map the input vector to high dimension space and using this space, it constructs the decision surface to ensure the generalized ability of the network. The main purpose of support vector machine to separate the training data without an error when it is impossible in this scenario. It must find the optimal hyperplane to separate the training data. Optimal hyperplane maximizes the margin between two classes. C parameter makes the best classification between two classes with the optimal hyperplane. More support vectors are required to separate the training data that optimize the margin between classes.

In another case, soft margin hyperplane is used when training data is not possible to separate without minimum error. So, training data can be separated with a decreased error. In soft margin analysis, to minimize the expense of error rate the C parameter is used with less value to separate the training data with minimum error. Sometime dataset have positive and negative instance overlapped with each other so it is difficult to classify the data. On the other hand, it may be over-fitted that cause computational complexity. This problem may be solved with C-SVM algorithms.

In support vector machine Coefficient C is used as a parameter that tolerates the systematic outlier in other class C-SVM tolerates less outlier in opponent classification. It holds a uniform value of C parameter for the positive and negative instances that help to satisfy of similar class distribution. Parameter C holds a value for positive and negative instance that satisfy the data set for distribution in classification. SVM interface depends upon the position of support vector. If a support vector found in opposition class then it influences the SVM interface, for this problem an error interface was built for a tolerance of support vector in opposite class. Value of parameter C allows less support-vector in the opposite class.

B. NU-SVM

NU-SVM is classification approach provided by Schölkopf et al. in [27]. It is used to control a large number of support vectors as well as training errors. Parameter v used upper bound and lower bound on the fraction of training errors and support vector respectively. Its range is between 0 and 1.

C. SMO

SMO stands for Sequential Minimal Optimization. The SMO method is proposed by Keerthi et al. [28]. This machine learning classifier used to train the Support Vector Machine. This new learning SVM learning algorithm is very simple, faster, easy to implement and having better-scaling properties. SMO perform well for sparse data set either it is binary or non-binary input data.

Sequential Minimal Optimization algorithm is used to solve the quadric programming problem. This algorithm decomposed the Quadric programming problems into sub-problems. It chooses the smallest optimization problem with two Lagrange multipliers to optimize jointly and find the optimal value for these multipliers. All Quadric programming problems solved quickly due to fast sub-problem.

SMO algorithm is best for avoiding extra use storage memory to store the 2 x 2 matrix. It solves the two Lagrange multipliers analytically. So a very large training problem can be solved in a personal computer that having less memory.

Three components for SMO like two Lagrange multipliers through the analytical method, a heuristic method for multiplier optimization and computing b method. For solving two Lagrange multiplier, this algorithm first computes the constraints and makes a solution for constrained maximum. Multiplier gives the name as script 1 and script 2 to multiplier 1 and 2 respectively that displayed on two-dimension. Constrained maximum lies on the diagonal line and this constraint explains why Lagrange multiplier is optimized.

Sequential Multiplier Optimization always maintains a feasible Lagrange multiplier vector. It increases overall objective function and converges asymptotically. SMO uses a heuristic approach to jointly optimize the Lagrange multipliers. One heuristic approach is for 1st Lagrange multiplier and one for 2nd Lagrange multiplier. The first heuristic approach provides an outer loop for 1st Lagrange multiplier that checks the overall objective function of the training set. If the first approach violates the KKT condition then check second multiplier KKT condition because it jointly optimizes the Lagrange multipliers.

V. Experimental Setup

A. Data Sets Description

Table I describes the biological munch mark databases used for the performance analysis of the decision tree based classifiers and SVMs in this empirical research study. Table I provide the information of data sets in terms of number of attributes, attribute type, number of instances and either missing values exist or not in the corresponding data set. The data sets are selected with significantly variant in database size, number of attributes and number of instances.

TABLE. I. DATA SET DESCRIPTION

Data Sets Name	No. of Attributes	Attributes Type	Missing Value	No. of instance
Lung-cancer	56	Integer	2	32
Lymphography	19	Categorical	None	148
Primary-Tumor	17	Categorical	N/A	339
Breast Cancer Dataset	9	Categorical	None	286
Dermatology	35	Categorical, Integer	Yes	366
Herbarman	3	Integer	None	165
Statlog	13	Categorical, Integer	None	270
Hepatitis	19	Categorical, Integer, Real	Yes	155

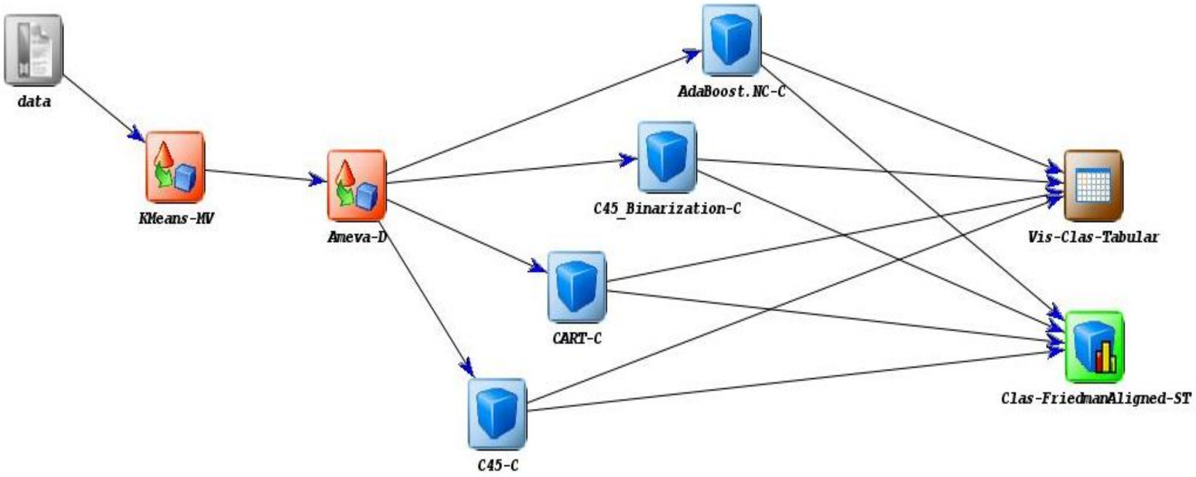

Fig. 1. Experimental Graph.

B. KEEL

Knowledge Extraction based on Evolutionary Learning (KEEL) is a data mining tool possessing various facilities for data preprocessing and different types of classification approaches for the comparison of new proposed classification approaches. It is a freeware java software tool. It provides a user-friendly GUI interface. It contains many built-in dataset and algorithm for data analysis. It provides many preprocessing techniques like feature selection, a method for missing value and hybrid models and statistical method for experiment.it use for educational and research purposes [8].

The current version of Keel has many advance features like multi-instance learning, subgroup discovery, semi-supervised learning and imbalanced classification. These features make versatility of the Keel improved and better deal with new data mining problems [29].

C. Experimental Graph

Fig. 1 shows the experimental graph generated in the KEEL. First stage data set loading, the second stage provide the facility of the imputation of missing values, the third stage provides the module for data discretization, the fourth stage shows the algorithms exploited the empirical study in this paper and final module provide the results of classifiers for the specific databases.

VI. RESULT AND DISCUSSION

This section provides the performance analysis of decision tree based classification approaches and support vector machines on medical databases in terms of accuracy and variance. Furthermore, the performance of a specific classifier is investigated in two fold; on a specific medical database and among the classification approaches.

A. Performance Analysis of Decision Tree based Classifiers

Table II shows the comparative performance analysis of AdaBoost.NC-C C4.5 –C, C4.5_Binarization–C and CART-C tree based classifiers that are chosen in this empirical research study. We compare the performance of these algorithms in Table II on different datasets in term of accuracy. The results show that C45-C and C45_Binarization-C provide equal accuracy on lung-cancer dataset. Moreover, C45-C also perform better on lymph, primary-tumor breast cancer dataset as compared to other algorithms in terms of accuracy. C45_Binarization provide the best performance in term of accuracy on Dermatology and Heart-statlog dataset. The AdaBoost.NC-C provide promising results on Hepatitis dataset; CART-C provides the best performance on Haberman dataset while the C45-C classifier provides 75.19% average accuracy on all datasets that is more promising comparatively w.r.t other classification algorithms. The C45-C_Binerization provide minimum accuracy of 6.06% and maximum accuracy 96.05 in percentage. Table III shows the comparative performance of the selected classifiers in terms of win/lose/draw. The win/lose/draw provides information, how many times a specific algorithm best performs to others.

From Table III, C45-C provides best accuracy on 4 selected datasets with respect to other classifiers. AdaBoost.NC and CART-C provide best accuracy only on one dataset and remaining 7 dataset loose by others algorithm. So AdaBoost.NC and C45_Binarization draw in one dataset.

The application of decision tree based classifier on selected dataset also provides performance in term of variance parallel. More variance on dataset provides lower performance result. CART-C provide bad performance on Lung-Cancer as well as on Dermatology and Hepatitis datasets as compared to AdaBoost.NC-C, C45_C, and C45_Binarization. C45-C provide variance on two selected dataset such as lymph and primary tumor and AdaBoost. NC-C classifier provides more variance on Breast cancer, Haberman, and Heart-Statlog. C45_Binarization –C classifier provides the best performance on selected dataset because there is no more variation as compared to other proposed classifier. CART-C provide 1.01% average variance and maximum 4.27% variance on selected dataset. C45-C provide minimum variance of 0.13% that is more than the other three classifiers. Fig. 2 provides more understandability of this decision tree based classifier's variance.

TABLE. II. DECISION TREE BASED CLASSIFIERS PERFORMANCE IN TERM OF ACCURACY (%)

Data Sets	**AdaBoost.NC-C**	**C45-C**	**C45_Binarization-C**	**CART-C**
lung-cancer	80.30	**83.33**	83.33	81.82
lymph	54.63	**78.10**	6.06	77.97
primary-tumor	19.56	**41.29**	37.79	35.93
breast-cancer	70.47	**74.64**	74.33	70.49
dermatology	45.91	94.05	**96.05**	53.92
haberman	66.51	71.83	71.83	**75.67**
heart-statlog	78.79	81.14	**81.48**	71.38
hepatitis	**80.08**	77.16	77.16	76.10
Average	62.03	**75.19**	66.00	67.91
Min	19.56	41.29	**6.06**	35.93
Max	80.30	94.05	**96.05**	81.82

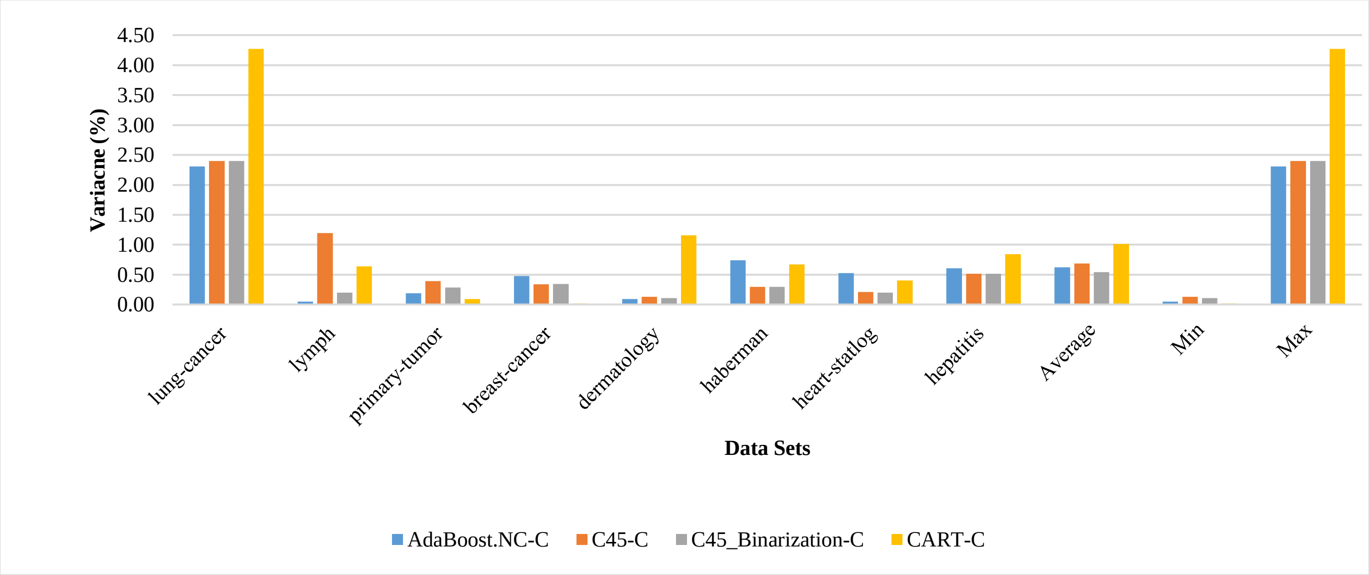

Fig. 2. Performance Analysis of Decision Tree based Classifiers in Term of Variance.

TABLE. III. STATUS COMPARISON OF DECISION TREE BASED CLASSIFIER

	Decision Tree Based Classifiers			
	AdaB.NC	**C45**	**C45_Bin**	**CART**
Win	1	**4**	2	1
Loose	7	**3**	5	7
Draw	0	1	1	0

B. *Performance Analysis of SVM based Classifiers*

Support Vector Machine performs classification tasks on the base of hyper-plane by using data point that is called support vectors. We used three support vector-based classifier on selected data set by using KEEL software. Table IV presents the results of comparative performance analysis of selected SVMs on corresponding medical databases. Support vector machine based classifier like SMO-C, NU_SVM-C and C_SVM-C are used in this proposed thesis on selected datasets. NU_SVM-C and C_SVM-C classifier provide best performance in term of accuracy on lung-cancer dataset. But C-SVM-C also provide best accuracy on primary tumor and breast cancer datasets 46.12% and 72.11% respectively. SMO-C provide best performance in term of accuracy on dermatology, haberman, heart-statlong, hepatitis and arrhythmia as compared to other two classifier but also proved average accuracy. C-SVM-C classifier provide minimum accuracy 46.12 and maximum accuracy 97.28 % accuracy as compared to other proposed SVM based classifiers. Table V provides the comparative performance of SVM based classifier in terms of win/lose/draw.

Table V shows the status of SVM based Classifiers with their performance for comparison. C_SVM-C classifier give best accuracy four time which is greater from other classifier. NU_SVM-C does not give best performance as compared to other even in one of the selected dataset. SMO-C gives performance in three classifiers.

Fig. 3 provides the comparative performance analysis of SVMs in terms of variance. SMO-C provide more variance on primary-tumor, Dermatology and hepatitis than NU_SVM-C and C_SVM-C and also provide minimum variance of selected variance as compared to other two classifiers. NU_SVM-C provide more variance on six datasets that make the performance bad on selected dataset as compared to other datasets. It also make more value of average variance on selected datasets that reach 1.26. NU_SVM-C and C_SVM-C provide equal maximum variance on selected datasets; as well as equal variance on lung-cancer dataset. All the information is highlighted in Table V.

Table VI provides the combined performance behavior of both categories Decision Tree-based classifiers and SVMs based classifiers in terms of accuracy. The performance of AdaBoost.NC-C classifier is lower than other methods on selected datasets. The C45-C provide the best performance on based of accuracy on lung-cancer and breast cancer datasets. C45_Binarization-C provide best accuracy result on Lung-cancer dataset equal to C45-C and Minimum average accuracy 6.06% as compared to remain six classifiers. CART-C provided the best performance on based of accuracy on Haberman dataset as compared to another dataset. Support Vector based algorithm SMO-C provided the best performance on based of accuracy on lymph, heat-statlog and hepatitis dataset and provided average accuracy as compared to other Decision tree and SVM based algorithms. NU_SVM-C accuracy is low to other both classifiers. C_SVM-C SVM based classifier provides the best performance on based of accuracy on primary tumor dermatology and provides maximum accuracy as compared to other classifiers on selected datasets.

TABLE. IV. PERFORMANCE ANALYSIS OF SVM BASE CLASSIFEIRS IN TERMS OF ACCURACY (%)

Data Set	**SMO**	**NU_SVM**	**C_SVM-C**
lung-cancer	70.45	**73.48**	**73.48**
lymph	**81.08**	71.90	75.54
primary-tumor	44.76	33.80	**46.12**
breast-cancer	69.25	61.95	**72.11**
dermatology	95.79	97.03	**97.28**
haberman	**75.09**	50.36	73.62
heart-statlog	**84.85**	73.06	82.49
hepatitis	**85.91**	81.78	85.38
arrhythmia	**62.03**	49.18	51.42
Average	**74.36**	65.84	73.05
Min	44.76	**33.80**	46.12
Max	95.79	97.03	**97.28**

TABLE. V. COMPARISION IN TERMS OF WIN/LOSE/DRAW

	Support Vector Machine Based Classifiers		
	SMO	**NU_SVM**	**C_SVM**
Win	**6**	0	4
Loose	3	7	3
Draw	0	1	1

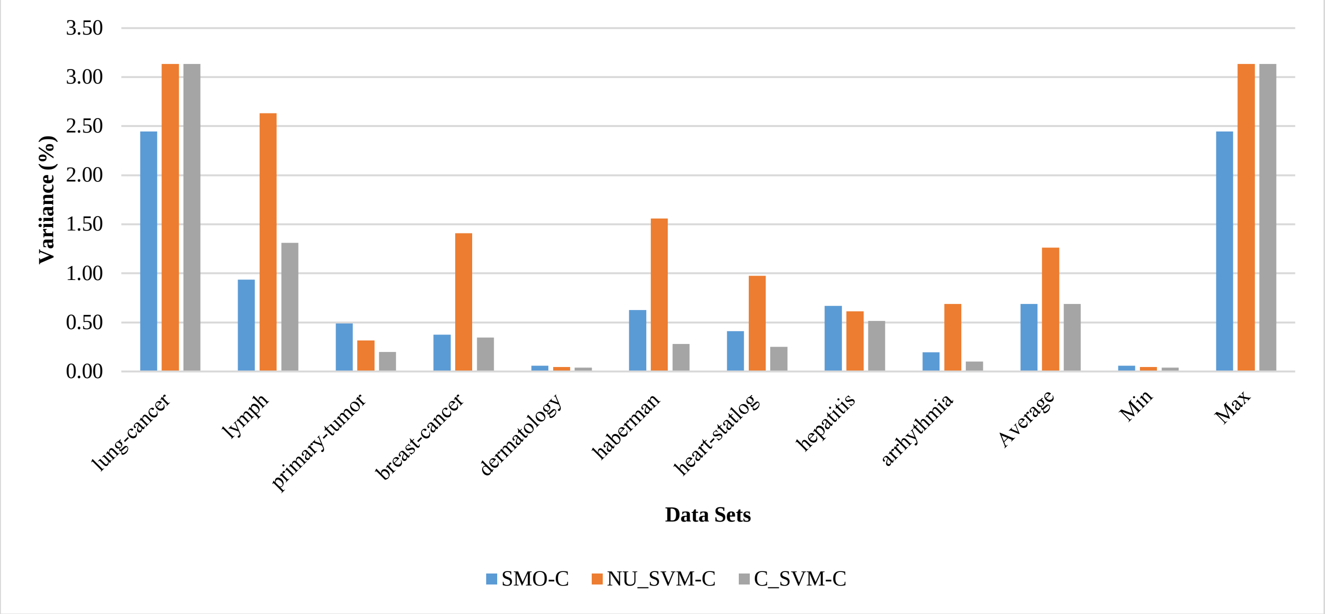

Fig. 3. Performance Analysis of Support Vector Machines in Term of Variance.

TABLE. VI. COMBINED RESULTS OF PROPOSED ALGORITHMS IN TERM OF ACCURACY

Data Sets	**Decision Tree Algorithms**				**Support Vector Machines**		
	AdaBoost.NC-C	**C45-C**	**C45_Binarization**	**CART-C**	**SMO-C**	**NU_SVM-C**	**C_SVM-C**
lung-cancer	80.30	**83.33**	83.33	81.82	70.45	73.48	73.48
lymph	54.63	78.10	6.06	77.97	**81.08**	71.90	75.54
primary-tumor	19.56	41.29	37.79	35.93	44.76	33.80	**46.12**
breast-cancer	70.47	**74.64**	74.33	70.49	69.25	61.95	72.11
dermatology	45.91	94.05	96.05	53.92	95.79	97.03	**97.28**
haberman	66.51	71.83	71.83	**75.67**	75.09	50.36	73.62
heart-statlog	78.79	81.14	81.48	71.38	**84.85**	73.06	82.49
hepatitis	80.08	77.16	77.16	76.10	**85.91**	81.78	85.38
Average	62.03	75.19	66.00	67.91	**75.90**	67.92	75.75
Min	19.56	41.29	6.06	35.93	44.76	33.80	46.12
Max	80.30	94.05	96.05	81.82	95.79	97.03	**97.28**

VII. CONCLUSION

Classification Rule Discovery from medical databases is a very hot and challenging problem in the field of Data Mining. There are several classification approaches proposed for the discovery of classification rules and prediction of diseases from medical databases. The choice of a classification method for the discovery of classification rules from specific medical databases still requires investigation of the suitability of classifiers in terms of performance analysis. This study investigates the performance of decision tree-based classifiers and Support Vector Machines on specific medical databases. The empirical performance analysis results reveal that C45-C performs better in terms of a total number of datasets while the overall average performance of C45_Binarization-C is better than other decision tree-based classifiers. The performance of SVM based classifiers, SMO-C is results are promising to NU_SVM-C and C_SVM-C in terms of accuracy. This research work provides the empirical performance analysis of decision tree-based classifiers and SVM on a specific dataset. Moreover, this paper provides a comparative performance analysis of classification approaches in terms of statistics.

In the future, this research work can be enhanced by increasing the number of medical databases with other statistical and evolutionary classifiers.

References

[1] O. Trelles, P. Prins, M. Snir, and R. C. Jansen, “Big data, but are we ready?,” Nature Reviews Genetics, vol. 12, no. 3, pp. 224, 2011.

[2] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework." pp. 82-88.

[3] D. A. Benson, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, “GenBank,” Nucleic acids research, vol. 42, no. D1, pp. D32-D37, 2013.

[4] T. W. Harris, J. Baran, T. Bieri, A. Cabunoc, J. Chan, W. J. Chen, P. Davis, J. Done, C. Grove, and K. Howe, “WormBase 2014: new views of curated biology,” Nucleic acids research, vol. 42, no. D1, pp. D789-D793, 2013.

[5] C. E. Bouton, A. Shaikhouni, N. V. Annetta, M. A. Bockbrader, D. A. Friedenberg, D. M. Nielson, G. Sharma, P. B. Sederberg, B. C. Glenn, and W. J. Mysiw, “Restoring cortical control of functional movement in a human with quadriplegia,” Nature, vol. 533, no. 7602, pp. 247, 2016.

[6] F. Thabtah, and D. Peebles, “A new machine learning model based on induction of rules for autism detection,” Health informatics journal, pp. 1460458218824711, 2019.

[7] R. Saravanan, and P. Sujatha, "A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification." pp. 945-949.

[8] J. Alcalá-Fdez, L. Sánchez, S. Garcia, M. J. del Jesus, S. Ventura, J. M. Garrell, J. Otero, C. Romero, J. Bacardit, and V. M. Rivas, “KEEL: a software tool to assess evolutionary algorithms for data mining problems,” Soft Computing, vol. 13, no. 3, pp. 307-318, 2009.

[9] Z. Iqbal, R. Ilyas, W. Shahzad, and I. Inayat, "A comparative study of machine learning techniques used in non-clinical systems for continuous healthcare of independent livings." pp. 406-411.

[10] S. O. Akinola, and O. J. Oyabugbe, “Accuracies and training times of data mining classification algorithms: An empirical comparative study,” Journal of software Engineering and Applications, vol. 8, no. 09, pp. 470, 2015.

[11] M. H. Tafish, and A. M. El-Halees, "Breast Cancer Severity Degree Predication Using Data Mining Techniques in the Gaza Strip." pp. 124-128.

[12] M. Ramasamy, S. Selvaraj, and M. Mayilvaganan, "An empirical analysis of decision tree algorithms: Modeling hepatitis data." pp. 1-4.

[13] Y. Yan, M. Chen, M.-L. Shyu, and S.-C. Chen, "Deep learning for imbalanced multimedia data classification." pp. 483-488.

[14] A. Anand, M. A. Haque, J. S. R. Alex, and N. Venkatesan, "Evaluation of Machine learning and Deep learning algorithms combined with dimentionality reduction techniques for classification of Parkinson’s Disease." pp. 342-347.

[15] S. Belarouci, F. Bekaddour, and M. A. Chikh, "A comparative study of medical data classification based on LS-SVM and metaheuristics approaches." pp. 548-553.

[16] S. Tharaha, and K. Rashika, "Hybrid artificial neural network and decision tree algorithm for disease recognition and prediction in human blood cells." pp. 1-5.

[17] D. A. Aljawad, E. Alqahtani, A.-K. Ghaidaa, N. Qamhan, N. Alghamdi, S. Alrashed, J. Alhiyafi, and S. O. Olatunji, "Breast cancer surgery survivability prediction using bayesian network and support vector machines." pp. 1-6.

[18] P. Hamsagayathri, and P. Sampath, "Priority based decision tree classifier for breast cancer detection." pp. 1-6.

[19] J. R. Quinlan, “Induction of decision trees,” Machine learning, vol. 1, no. 1, pp. 81-106, 1986.

[20] J. Quinlan, “C4. 5: Programs for machine learning. Morgan Kaufmann, San Francisco,” C4. 5: Programs for machine learning. Morgan Kaufmann, San Francisco., pp. -, 1993.

[21] S. Wang, and X. Yao, “Multiclass imbalance problems: Analysis and potential solutions,” IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, no. 4, pp. 1119-1130, 2012.

[22] L. Breiman, J. Friedman, R. Olshen, and C. Stone, “Classification and regression trees (Chapman y Hall, Eds.),” Monterey, CA, EE. UU.: Wadsworth International Group, 1984.

[23] R. Gholami, and N. Fakhari, "Support Vector Machine: Principles, Parameters, and Applications," Handbook of Neural Computation, pp. 515-535: Elsevier, 2017.

[24] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C.-C. Chang, C.-C. Lin, and M. D. Meyer, “Package ‘e1071’,” The R Journal, 2019.

[25] R. Gandhi, “Support Vector Machine—Introduction to Machine Learning Algorithms,” Towards Data Science, 2018.

[26] C. Cortes, and V. Vapnik, “Support-vector networks,” Machine learning, vol. 20, no. 3, pp. 273-297, 1995.

[27] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, “New support vector algorithms,” Neural computation, vol. 12, no. 5, pp. 1207-1245, 2000.

[28] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, “Improvements to Platt's SMO algorithm for SVM classifier design,” Neural computation, vol. 13, no. 3, pp. 637-649, 2001.

[29] I. Triguero, S. González, J. M. Moyano, S. García López, J. Alcalá Fernández, J. Luengo Martín, A. Fernández Hilario, J. Díaz, L. Sánchez, and F. Herrera, “KEEL 3.0: an open source software for multi-stage analysis in data mining,” 2017.