Optimised Heart Disease Prediction Scheme

Dr. L. Pavithira[1], Dr. Wilfred Blessing N R[2], Dr. I Shanmugapriya[3], Dr. SK Wasim Haidar[4], Sutherlin Subitha G[5]

1Associate Professor, Department of Computer Science, PSG College of Arts &Science College, Coimbatore, Tamilnadu, India

2Senior Lecturer, IT, College of Computing and Information Sciences, University of Technology and Applied Sciences – Ibri, Sultanate of Oman.

3Assistant Professor, Department of Computer Science, PSG College of Arts and Science, Coimbatore, Tamilnadu, India.

4Lecturer, IT, College of Computing and Information Sciences, University of Technology and Applied Sciences – Salalah, Sultanate of Oman.

5Assistant Professor, Department of CSE, Stella Mary's College of Engineering, Kanyakumari District, Tamilnadu, India.

Abstract:

Cardio Vascular Disease (CVD) includes heart, peripheral arterial and coronary heart disease as well as cerebrovascular disease. The symptoms of this disease include fast heartbeats, pain or discomfort in the centre of the chest, dizziness, difficult breathing, overweight, hypertension and high cholesterol. To determine accurate and efficient treatment of heart disease, specialized cardiologists must conduct complicated tests and procedures. Early detection is the key to preventing CVD. Difficulties in diagnosing CVD and transporting patients over long distances lead to an unjustified increase in death rate which is a burden on patients. Among important risk factors for cardiovascular problems are an abnormal build-up of fats and cholesterol in the blood vessels, harmful alcohol consumption, smoking and a lack of proactive steps, and poor dietary habits. Early detection and cardiovascular management are essential factors to lowering the incidence of CVD. Applying an optimized technique would be a better choice for such classification. The proposed system uses Modified Teaching Learning Based Optimization (MTLBO), Kernel Density Function (KDF) and Density-based Modified Teaching Learning Based Optimization (DMTLBO) to perform dataset classification. Classification is performed using Support Vector Machine (SVM), Ensemble Learning (EL-Adaboosting). It is seen that DMTLBO_Adaboosting offers better results based on Accuracy, Precision, Recall, Time Period, F-Measure as well as Error Rate.

Keywords: Cardio Vascular Disease (CVD), Support Vector Machine (SVM), Ensemble Learning (EL), Teaching-Learning-Based Optimization (TLBO), Modified Teaching-Learning-Based Optimization (MTLBO), Kernel Density Function (KDF), Density based Modified Teaching Learning Based Optimization (DMTLBO)

DOI: 10.24297/j.cims.2024.8.1

1. Introduction

Cardio Vascular Disease (CVD) is the main cause of mortality. It can be prevented to a greater extent by early diagnosis [1]. Accurate detection and diagnosis of disorders is a key challenge with medical data analysis. Healthcare gathers remarkable information that includes data regarding coronary disease diagnosis which is involved on accurate decision [2]. CVD is a wide term for condition of attacking blood vessels of heart owing to excess accumulation of fatty deposits in arteries called atherosclerosis [3]. It leads to a high risk of blood clot formation in blood vessels that blocks blood circulation.

CVD is a family disorder that contains inherited coronary symptoms, peripheral vascular disease and coronary heart cerebrovascular illness [4]. Each four out of five CVD cases die owing to coronary artery disease and cerebrovascular accidents based on report of World Health Organization (WHO) [5]. Based on WHO, CVD is the main reason of death globally. CVD causes failure in human heart and blood vessels. It involves several symptoms like fast heart rate, chest pain, dizziness, difficulty in breathing, obesity in increased blood pressure and cholesterol and so on [6, 7].

Identifying heart disease demands expert cardiologists with complex procedures as well as tests to figure out precise and effective treatment [8, 9]. CVD may be diagnosed at a later stage or patients may be transported over long distances [10]. The critical causes that increase the probability of CV disorders are abnormal accumulation of fats and cholesterol in blood vessels, hazardous use of alcohol, smoking and absence of proactive tasks, hypertension, high sugar level and bad eating habits [11, 12]. The way to decrease the rate of CVD lies in the consideration of early prediction and CV management.

2. Related Work

The works done by numerous authors associated with feature extraction and classification of lung cancer images are detailed below.

2.1 Feature Extraction

Machine Learning (ML) based techniques are applied on CVD datasets. They offer different results. Classification precision and selection of vital attributes are based on effectiveness of medicinal analysis. El-Bialy et al. (2015) [13] have incorporated ML analysis implemented on CVD datasets. This overcomes the challenges related to misplaced, inappropriate and unreliable data that may occur during data collection. Decision Tree (DT) as well as pruned C4.5 tree is employed, as resultant trees are mined from dissimilar datasets and compared. Common attributes between the datasets are mined and considered in future study for similar ailment in any dataset. Classification accuracy of is 78.06% more than that of discrete datasets which is 75.48%.

Long et al. (2015) [15] have proposed a detection system using rough set centred feature lessening and Interval Type-2 Fuzzy Logic System (IT2FLS). The combination among these two approaches target to deal with challenge and uncertainties related to high-volume input. IT2FLS exploits a fusion learning procedure including Fuzzy C-Means clustering procedure and factor tuning by chaos firefly and genetic fusion procedures. The learning method involves more amounts of computations, particularly when engaged with dataset involving more dimensions. Rough set centred feature reduction based on chaos firefly procedure is examined to discover optimum decrease that consequently decreases computational weight and improves performance of IT2FLS. Experimentation outcomes determine significant supremacy of proposed scheme in contrast to ML approaches like Naive Bayes (NB), SVM and Artificial Neural Network (ANN). This acts as a decision support system for heart ailment analysis.

Chavan &Sonawane (2017) [15] have developed a low cost treatment using data mining tools for enabling data based decision support system. The analysis of heart sickness using dissimilar attributes or indications is complicated. Two DM classification methods like ANN and NB are considered to support in the analysis of the heart ailment and treatment. It is essential to screen numerous medicinal factors and post-operational days. Therefore, the modern trend in medical sector makes use of IoT in recent times. AVR-328 microcontroller is taken as an entry to link sensors for monitoring drip stages and observe the movement. The micro-controller takes data from sensors and directs it to the system over Wi-Fi and offers real-time observation of healthcare factors for clinicians. The records can be retrieved anytime by the clinician. The controller is also associated with beeper to alert the caretaker about deviation in sensor output. During extreme situation, alert note is directed to the surgeon over the android app attached to the cloud server. So, rapid provisional medication can be certainly done and patient’s details can be found automatically. This scheme is effective with little power usage, informal setup, high performance and prompts response.

DM converts the huge group of fresh healthcare content into sensible information that can support enhanced decision and forecast. There are certain studies that use DM methods in heart disease forecast. However, studies that focus on important attributes for forecasting CVD are restricted. It is vital to choose the right grouping of important attributes that can improve the presentation of the forecast models. A min et al. (2019) [16] have aimed to find the important attributes and DM techniques that ensure correctness. Prediction models are improved using various combinations of attributes, and 7 classification methods: K-Nearest Neighbour (KNN), Decision Tree (DT), NB, SVM, Logistic Regression (LR), NN and Vote. Research outcomes confirm that heart disease forecast model established using recognized important attributes and best-performing DM method achieves an accuracy of 87.4%.

Different attribute determination schemes are available to handle data with multiple features. These attributes support data by using extremely high measurement values. The attribute determination approach offers a technique to decrease execution time, improve forecast execution, and also provides enhanced consideration of data in ML as well as a mode to identify applications. Present works emphasis on increasing classification precision. In case of real-world applications, the designated subsection of attributes should be constant instead. Aggrawal & Pal (2020) [17] have proposed a progressive attribute selection procedure to identify deaths due to heart problems in the course of treatment. Also, the precision of this technique is associated with classifier accuracy. The confusion matrix, recall, ROC curve, F1-score and precision are computed to confirm outcomes of this approach. The investigational results indicate that for Random Forest (RF) based feature selection, Sequential Feature Selection (SFS)offers 86.67% precision.

Researchers have proposed different intelligent investigative systems to diagnose the cardiac problem. Reduced precision in heart disease prediction is challenging. For enhanced heart risk forecast precision, Javeed et al (2020) [18]have suggested an attribute selection technique that makes use of moving window with adaptive dimension for extraction of attributes. Followed by this attribute removal process, 2 types of classification frameworks like ANN along with Deep Neural Network (DNN) are employed. Two categories of hybrid diagnostic systems like Floating Window with Adaptive size for Feature Elimination (FWAFE) ANN and DNN are involved. The efficiency of the suggested approaches is assessed on a dataset gathered from Cleveland online heart ailment dataset. The advantage of the suggested approach is evaluated in terms of Matthews Correlation Coefficient (MCC), precision, sensitivity, specificity and Receiver Operating Characteristics (ROC) curve. Investigational results authorize that recommended models outperform several approaches in the past that achieve accuracies in series of 50.00 - 91.83%. Furthermore, performance of projected models is extraordinary as equated with other advanced ML methods for heart ailment analysis. Also, the suggested schemes can support Physicians to obtain precise results.

Absence of exact analysis of heart disease is a major concern but there are certain restrictions to overcome this. An automatic investigative method for medical heart disease prediction is offered by Shah et al. (2020) [19]. They have computed the most appropriate attribute subset by considering the benefit of attribute selection and mining procedures. To complete attribute selection process, two procedures namely, Mean Fisher centred and accuracy based attribute chosen procedure are presented. The designated attribute subset is then further advanced over attribute mining technique. The recommended scheme is authenticated over Cleveland, Hungarian, Switzerland and incorporation of their combination. Radial Basis Function (RBF) kernel-based SVM isemployed for categorising people with heart disease. The recommended approach is assessed based on specificity, sensitivity and accuracy metrics.

A prediction system for CAD is proposed by Valarmathi & Sheela (2021) [20]. Three Hyper Parameter Optimization (HPO) methods like Grid search, genetic programming and randomized search are suggested to enhance presentation of RF and XG Boost classifier model. The results of these classifiers are associated with prevailing studies. The performance of models is assessed with openly obtainable datasets namely, Cleveland Heart disease Dataset and Z-Alizadeh Sani dataset. The classifiers offer increased precision of 98% for CHD Dataset. RF with arbitrary search offers the maximum precision of 80%, 74% and 77% for analysis of three levels using Z-Alizadeh Sani Dataset. The outcomes are compared with present studies.

ML models may have specific intrinsic difficulties like efficient attribute selection, splitting features and unfair forecast of datasets. Many bulk datasets include multi-class tags, but groupings are in dissimilar extents. Magesh & Swarnalatha (2021) [21]have taken Cleveland's input from UCI source and recommended a decision tree based on cluster tree. This primarily comprises of five crucial phases. Initially, the original set is divided over target label dissemination. From high distribution, probable class combination is made. For every class-set arrangement, important attributes are recognized over entropy. With the important critical attributes, entropy -centred division is carried out. Finally, on these entropy groups, performance of RF is made by means of selected features in forecasting heart ailment. From this method, the RF classifier attains 89% better prediction accuracy from 77%. Therefore, the fault rate of RF has considerably lessened from 23% to 10%.

2.2 Classification

Classification, which is a kind of supervised ML enables estimation for future cases centred on a former dataset. A brief description of the most extensively used classification approaches for prediction of heart ailment is given below.

Two familiar algorithms utilized by Bhuvaneswari & Kalaiselvi (2012) [22]in DM classification include Back propagation Neural Network (BNN) and Naive Bayes (NB). Bayesian methods are basically imperative DM method. Given probability distribution, Bayes classifier can provably attain optimum outcome. This is centred on likelihood theory. Bayes rule is considered to compute subsequent from previous classification as latter is typically easier to be designed from a probability model. Statistics offer a solid background for quantification and estimation of outcomes. Though, procedures centred on data need to be revised and scaled before they are applied to DM.

Genetic Algorithm (GA) is an evolutionary procedure that is centred on Darwin's concept of progress. It reproduces approaches in nature like mutation, crossover, and natural selection. Waghulde & Patil (2014) [23] has highlighted the important benefit which is its practice to set loads of NN model. Usage of ANN along with GA is observed in numerous studies to create a hybrid prediction model.

ANN as indicated by Pouriyeh et al. (2017)[24] was established to reproduce the nerves in the human mind. It comprises of certain vertices or nerves that are associated and the outcome of one vertex is the input of another. Every vertex accepts numerous inputs, but the output is a single value. MLP is an extensively utilized category of ANN, and it comprises of input, hidden and output layers. Dissimilar amount of neurons are allocated to every layer in various situations. RBFis a kind of ANN is comparable to the MLP-NN but has a dissimilar quantity of hidden layers, approximation procedures, amount of factors, & additional features.

They have acclaimed that DT has flowchart-like layout. It comprises of branches, leaves, vertices and a root node. The inner vertices comprise the features, while branches signify outcome of every test on every node. It is extensively taken for categorization of resolutions as it doesn't require abundant awareness or setting factors. Authors have endorsed that K-Nearest Neighbour (KNN) forecasts the class of novel illustration centred on the votes by its nearby neighbours. It takes Euclidean distance to compute the space of a feature from its nearby values.
SVM has valuable classification exactness. The authors have defined it as a finite vector space that comprises of 1-dimension for every object feature.

The authors have also claimed that NB classifier fits to a cluster of probabilistic classifiers centred on NB theorem. It accepts robust independence among attributes, and this is the important portion of how this classifier makes estimates. It is easy to construct, and it typically achieves good results that makes it appropriate for analysis.

Ensemble Learning (EL) projected by the authors integrates several classifiers into single model to upturn precision. There are 3 kinds of EL techniques. Type 1 is bagging that is combining classifiers of same kind by voting mechanism. The 2nd kind is boosting that is similar to bagging. However, the new model influences the outcomes of previous models. Stacking is third type which means combining ML classifiers for many categories to create a model.

Taking the difficulty of arrhythmia discovery, an enhanced Convolution Neural Network (CNN) model is presented for precise classification. In contrast to conventional ML methods, CNN needs no extra attribute mining procedures because of availability of automatic attribute computing layers. An enhanced CNN is proposed by Wang et al. (2020) [25]to spontaneously categorize heartbeat of arrhythmia. Initially, heartbeats are separated from original signals. After separation, ECG heartbeats are entered into 1st convolution layers. Kernels with various sizes are involved in every convolution layer that considers advantage of attributes in diverse measures. The outcomes of max-pooling layer are combined and passed to fully-connected layers. The experiment is aligned with AAMI inter-patient standard that involves regular beats, supra-ventricular and ventricular ectopic beats, fusion & unknown beats. For authentication, MIT arrhythmia record is presented to authorize accurateness of suggested technique, and comparative experimentations are conducted. An accuracy of 99.06% is obtained. Incorrectly trained, the enhanced CNN model maybe spontaneously used to identify diverse kinds of arrhythmia from ECG.

3. Feature Extraction

In this work, Teaching-Learning-Based Optimization (TLBO), Modified Teaching-Learning-Based Optimization (MTLBO), Kernel Density Function (KDF) and Density based Modified Teaching Learning Based Optimization (DMTLBO) are used for efficiently extracting features.

3.1 Teaching Learning Based Optimization (TLBO)

Optimization is a process of generating optimal solutions for all parameters of an object function in terms of cost and efficiency. TLBO is a meta-heuristic algorithm and is a population-based optimization scheme that reproduces environment of classroom or population with a fixed number of students as learners. Where the teacher interacts with students and delivers knowledge to students, and teacher makes learners of a class educated. Learners can gain and improve their knowledge by interacting with teacher and themselves.

TLBO incorporates two phases of learning, one is teacher phase, another is a learner stage. It is a general computation method, where integration of understudies carried out as a general population for various problems proposed to understudies corresponding to different kinds of optimization problems. Eventual outcomes of every student are equivalent to non-diseased rate of optimization. The best outcome in the overall population can be considered as an educator. Implementation and performance of TLBO methodology are clearly defined in the teacher-student phase.

Teacher phase

In this stage, the teacher analyses the performance of students by sharing information among students. Consider ‘n’ subjects allotted for ‘m’ learners. Assume a population size of ‘K’ varies from 1 to n. Since teacher is the most skilled person and capable individual with respect to the subject and the best student in whole population is taken as an educator. The eventual outcome of the best performer with respect to all the subjects can be considered as the teacher by calculating ‘X_total- k_(best,i)’. The teacher gives best effort share the knowledge to the entire class eventually.

Students get data as demonstrated by the way of education provided by teacher and nature of students. Pondering to this reality, qualification among eventual outcome of educator and mean outcome of student in every subject is imparted as:

$\mathrm{Diff\_mean}_{j,i} = r_i(X_{j,k_{best,i}}, T_{fM_{j,i}})$(1)

Where,

$X_{j,k_{best,i}}$ - Consequence of teacher in subject ‘j’

r_i - Random number in the range [0, 1]

Teaching Factor (TF) picks the mean assessment to be altered. The assessment of TF can be either ‘1’ or ‘2’. It is picked discretionarily with identical probability as:

$\mathrm{TF} = \mathrm{round}[1 + \mathrm{rand}(0,1)\{2-1\}]$(2)

Where,

rand - Irregular number in reach [0, 1]

TF is not a boundary for this computation. Assessment of TF cannot be considered as the computation and its value is hardly picked by the estimation using the following Equation. The current course of action is revived in the teacher stage according to the going with explanation:

$X^{i}_{j,k,i} = X_{j,k,i} + [\mathrm{Diff\_mean}_{j,i}]$(3)

Where,

$X^{i}_{j,k,i}$-Newcapacity of '$X_{j,k,i}$'

Where,

'$X^{i}_{j,k,i}$' -Unrivalled limit regard

The recognized limits regards at the completion of the teacher stage are defined, and these attributes are set to be the estimated outcome to student level. It is clearly seen that the assessments of 'r_i' and TF denotes Teaching Factor which impacts introduction of TLBO estimation. 'r_i' is the self-assertive factor in reach [0, 1]. Nevertheless, assessments of 'r_i' and TF are determined arbitrarily and these boundaries are not given as commitment to the computation.

'r_i' and TF need not be tuned in TLBO estimation. Standard control boundaries like people size and ages etc., are to be tuned and ordinary control boundaries are essential for population based on enhancement computations. Thus, TLBO is known as a computation unequivocal boundary-less estimation.

Student stage

Student or learner phases of evaluation regenerate the performance of students through participation among themselves. Data can be acquired by sharing and helping various students. An understudy adjusts the information assuming various understudies having more amounts of data than the individual being referred to. Arbitrarily choose two students, 'A' and 'B'.

$X'_{total-A,i} = X'_{total-B,i}$ (4)

where,

$X'_{total-A,i}$ - Fresh estimation of $X_{total-A,i}$

$X'_{total-B,i}$ - Fresh estimation of $X_{total-B,i}$

At the end of educator stage,

If $X'_{total-p,i} > X'_{total-Q,i}$,

$X''_{j,p,i} = X'_{j,p,i} + r_i(X'_{j,p,i} - X'_{j,Q,i})$

If $X'_{total-Q,i} > X'_{total-p,i}$,

$X''_{j,p,i} = X'_{j,p,i} + r_i(X'_{j,Q,i} - X'_{j,p,i})$

'$X''_{j,p,i}$' is accepted if it gives prevalent limit.

3.2 Modified Teaching Learning Based Optimization (MTLBO)

Modified Teaching Learning Based Optimization (MTLBO) is used for classification of coronary disease. The two stages repeatedly functions until the condition is satisfied. The number of iterations depends on problem statement and criteria given for classification. The Best students are the finest outcomes in a specific cycle and interaction can be done between the students irrespective of the attributes. Further development of TLBO for a solitary cycle is as per the following.

Teacher Phase

The instructor stage is defined for 't' number of occurrences.

Stage 1: Initialize various occurrences and split the population into 2 sets (g) depending on classes, i.e., dangerous and not dangerous from data. Characterize the number of features.

Stage 2: Compute the mean of every selected features for all elements.

Stage 3: Find the best person in every set depends on their physical condition. The wellness of every individual will be determined regarding each person in the other set and select the best result as educators from each set. SVM and EL algorithms are utilized to track down the precision of every element and pick the best value. Generally, the educator is considered as a very scholarly person who instructs students to work on their outcomes.

Step 4: Analyse the difference mean for every one of features through best value. It can be assessed as reduction of features from the best particle. TF is a function represented by TF with values 1 or 2 for the best particle and it is the random number that varies from 0 to 1. The best individual is chosen as the educator and trains students to further enhance their performance depending on the capacity.

Stage 5: If the result is identified as best to the result identified with the expected outcome, substitute the earlier value.

Learner Phase

Analysis of learner stage regenerates the performance of students through participation among themselves. Data can be acquired by sharing and helping various students. Learner phase can be executed by selecting two instances 'X' and 'Y' randomly, as they are restructured attributes. If outcome is better than the expected value, then consider the best value as teacher. Otherwise, continue previous values. If the end condition is satisfied, then submit the final outcome or repeat Step 2. The population includes only particles with ideal values and a set of attributes after more number of iterations for educator and student phases. For improving performance, new data should be trained by the classifier.

The principal outcome of the proposed system is:

The educator class includes multiple instructors from different generations to investigate the entire feature space and handle local optimization problems. It causes new system to merge with normal speed.

Instead of using random number as teaching factor, direct values of 1 or 2 can be applied. It permits algorithm to use the solution for a universal maximum.

The method of modifying solutions is enhanced in two phases of TLBO by choosing random values.

```
MTLBO Algorithm
For i =1 to D
    A=Select a random integer {1,2,….N}
    $M_i = X_i^a$
End
For (k = 1 to N)
    If ($r_1$< 0.5)
        For (i = 1 to D)
            If ($r_2$< P)
                $T_i^f = \text{round}\ ( 1 + r_2 )$
                $X_i^{new} = X_i^k + r_3(X_i^{best} - T_i^f \times M_i)$;
        End For
    Else
        For (I = 1 to D)
            If(i< P)
                R =select a random integer in
                {1,2………N}
                If ($X^r$ is better than $X^k$)
                    $X_i^{new} = X_i^k + r_4 (X_i^r - X_i^k)$
            Else
                $X_i^{new} = X_i^k + r_5\ (X_i^{r\square} - X_i^k)$
            End if
          End if
      End For
End
If ($X^{new}$is better than $X^k$ )
                              $X^k = X^{new}$
End
```

Where,

‘r_1’, ‘r_2’, ‘r_3’, ‘r_4’ and ‘r_5’ - Uniformly distributed random numbers between [0, 1]

3.3 Kernel Density Function (KDF)

Kernel Density Function (KDF) is non-parametric and it doesn’t build any agreement concerning data distribution. It generally picks features that catch the presentation of common information by isolating the anomalies. A forward search technique is utilized for standard estimation. This is profoundly equipped for finding anomalies in contrast to other techniques. Because of parallelism, integration of different searching algorithms becomes more challenging related to feature selection.

3.4 Density based Modified Teaching Learning Based Optimization (DMTLBO)

DMTLBO is employed to simplify the general TLBO in computation of evaluating function. Dimensions of the attribute and input size are examined to be parameters to obtain biased group of attributes.

It initiates by defining the population scope ‘t’ and design variable ‘s’ that is ‘n’ quantity of learners and ‘m’ amount of subjects who are trained.

The objective function is computed as follows:

Min f(y)=∑nr=[y.2r –10 cos(2πyr)+10](5)

Teacher phase

The ideal teacher is selected in this stage. The teacher gives his fullest effort to share knowledge with whole class eventually. Students get data as demonstrated by the way of education provided by teacher and nature of students. The last iteration value may be determined as ‘y^th’ iteration for (y=1, 2, 3....m)for x(x=1, 2, 3.....n). The mean of individual subject is taken and illustrated as ms(x,y).

Variances are considered to modernize standards in resolution pool by consolidating differences to current solution and method involved in the learner stage. The Chebyshev Distance (D_c) is used to improve values in output function. Difference is represented as 'D_s'.

'D_s'.

$D_s = v\,(O_{new,s} - TFO_s)$ (6)$D_c(y_i, y_j) = \max\,(|y_i - y_j|)$ (7)

$X'_{new} = f(y) + D_c(y_i - y_j)$ (8)

Learner phase

By interacting with student's group, the knowledge of each student may be enhanced.

Algorithm:

```
for $y = 1: t_r$
    Choose additional learner arbitrarily$X_x$, $y \neq x$
        If $f(X_X < f(X_X)$
            $X''_{new,y} = X'_{new,y} + r_y(X_Y - X_X)$
        Else
            $X''_{new,y} = X'_{new,y} + r_y(X_X - X_Y)$
        End If
End For
```

Declare 'X_{new}', if function value is greater than previous value. The features that show improved results depend on most recent evaluation methods through every cycle are collected in the subset of attributes. This method ends when every attribute is taken for analysis.

4. Classification

In this work, SVM and EL are used for classification.

4.1 Support Vector Machine (SVM)

It is employed in classification that addresses training data as focused in space that is isolated into hyperplanes. The objective of classification is to accurately forecast the objective class for every condition in insights. These systems are verified by differentiating the outcome with predicted values. The essential thing in classification is partition of important and unimportant attributes and forecast of accurate class name.

To overcome this issue, SVM is introduced in the propounded system for CVD characterization. It isolates the classes by hyper-plane and uncovers superior execution of applications. The principal motive is to amplify the gap between hyperplanes and neighbour nodes to reduce error rate. The accuracy of work is analysed by 3 variations of SVM classifier like P-SVM, R-SVM and L-SVM taken from the LIBSVM libraries in the propounded system. There are 'n' samples in datasets, and assumed training set is limited.

4.2 Ensemble Learning

Adaptive Boosting (adaboost) classifier is a boosting method and it is combined with numerous linear weak classification algorithms. In feature selection, every weak classifier depends on an aspect of the classification. It is a technique that converts weak understudies into strong ones by selecting the best classifier and further develops the model by training incorrectly classified features recursively from pre-defined dataset. The system can self-adaptively build the number of powerless classifiers for further enhancement in terms of accuracy of system and focus on principal features. The classification utilizes the lowest error value toimplement weak classifier weights and rearrange the value of every training model after adding weak classifiers, and the results are substituted to weak classifier. The impact of general parallel classifier can be enhanced. EL algorithm is a solid classification algorithm made out of numerous weak classifiers.

5. Results And Discussion

Chronic Kidney Disease (CKD) dataset and Wisconsin Diagnostic Breast Cancer (WDBC) are taken from UCI ML repository. CKD dataset includes 25 attributes and 400 instances and WDBC dataset comprises of 32 attributes along with 569 Instances.

5.1 Chronic Kidney Disease (CKD) Dataset

CKD dataset comprises of tests and several measures taken from patients. The details are composed from 400 patients in observation for 60 days. 250 are detected with CKD and 150 are without CKD. This disparity is signified as 'Class'. Some significant features of the dataset are age, Hypertension, Diabetic, Blood Urea, Haemoglobin, Blood Glucose Random etc. Table 1 shows the number of features using TLBO, KDF and DMTLBO for CKD dataset.

Table 1: Amount of Features of CKD Dataset

No. of Attributes	TLO	KD	DMTLO
25	18	16	19

Table 2 shows the features selected using TLBO, KDF and DMTLBO for CKD dataset. Table 3 shows the performance for CKD dataset.

Table 2: Features of CKD Dataset

TLO	KD	DMTLO
2,3,4,5,6,10,17,18,19,14,15,11,13,12,9,8,16,20	3,4,5,10,11,12,15,16,19,18,11,8,9,2,14,6	2,3,10,4,5,17,18,19,14,15,6,7,11,12,13,8,9,21,23

Table 3: Performance of CKD Dataset

Algorithm	Accuracy	Precision	Recall	F-Measure	Time Period	Error Rate
MTLBO_SVM	91	89	91	91	3.8	21
MTLBO Adaboosting	94	92	93	93	3.2	19
KDF_SVM	90	88	89	90	4.3	22
KDF_Adaboosting	93	90	92	93	3.7	21
DMTLBO_SVM	95	93	94	95	2.9	16
DMTLBO_Adaboosting	97	95	96	96	2.3	12

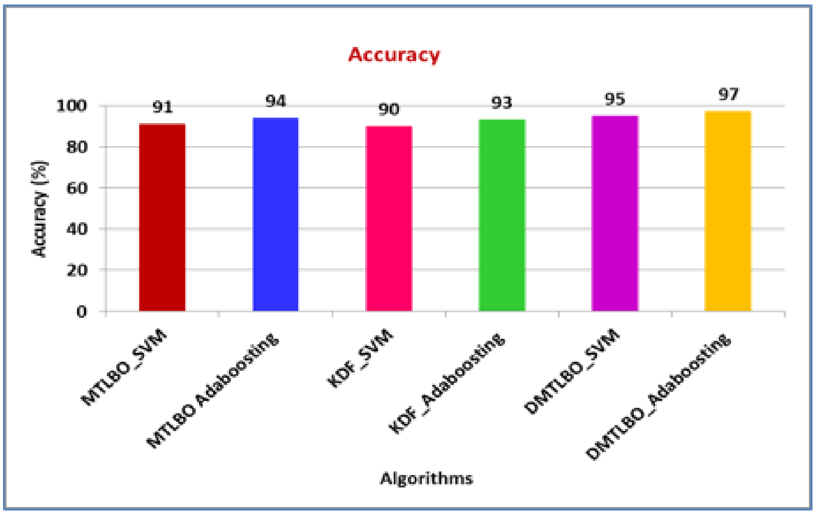

Figure 1: Accuracy

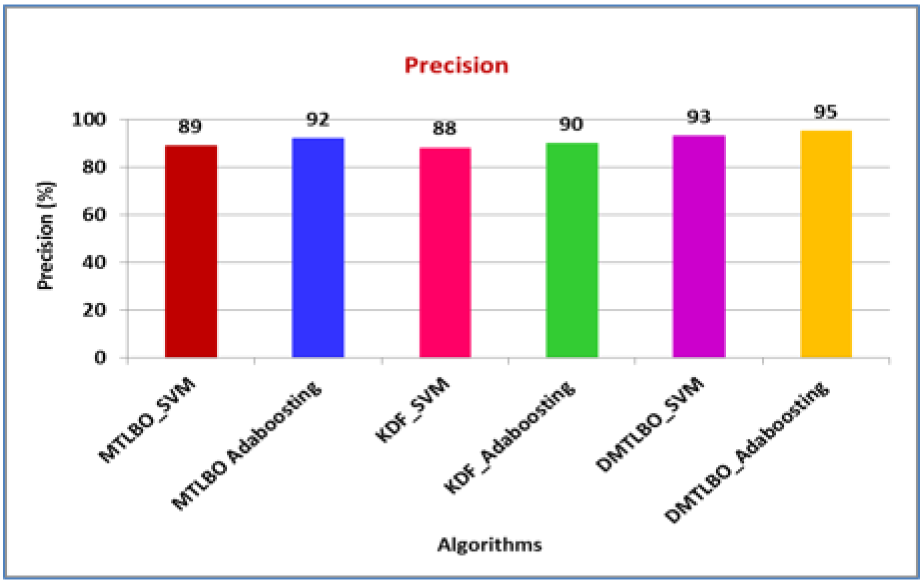

Figure 2: Precision

Figure 1 gives the Accuracies of MTLBO_SVM, MTLBO Adaboosting, KDF_SVM, KDF_Adaboosting, DMTLBO_SVM and DMTLBO_Adaboosting schemes. The propounded DMTLBO_Adaboosting scheme offers 6%, 3%, 7%, 4% and 2% improved Accuracy when compared to the benchmarked schemes. Figure 2 shows the Precision of the benchmarked schemes. The propounded DMTLBO_Adaboosting scheme offers 6%, 3%, 7%, 5% and 2% enhanced Precision when compared to the standard schemes.

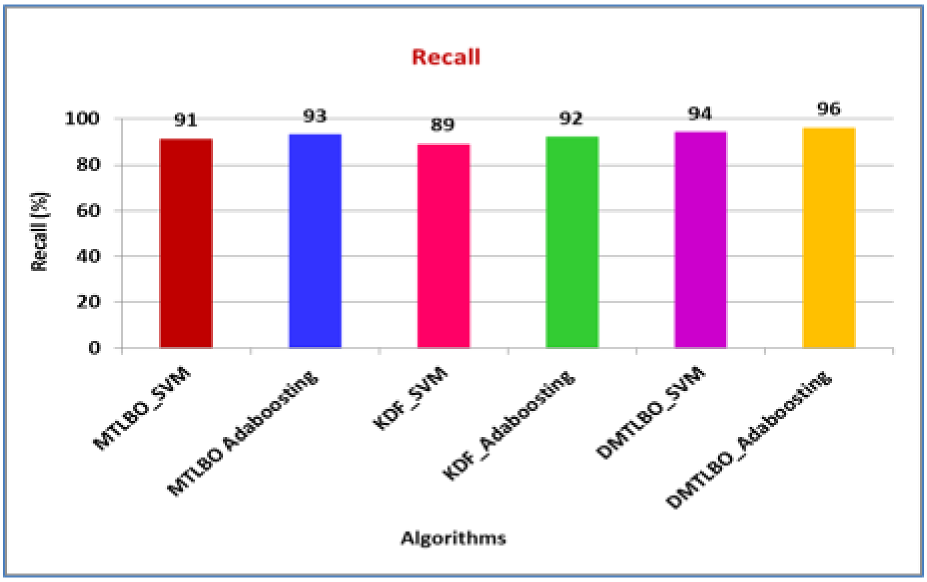

Figure 3: Recall

Figure 3 presents the Recall of the benchmarked schemes. The propounded DMTLBO_Adaboosting scheme yields 5%, 3%, 7%, 4% and 2% enhanced Recall in contrast to the standard schemes.

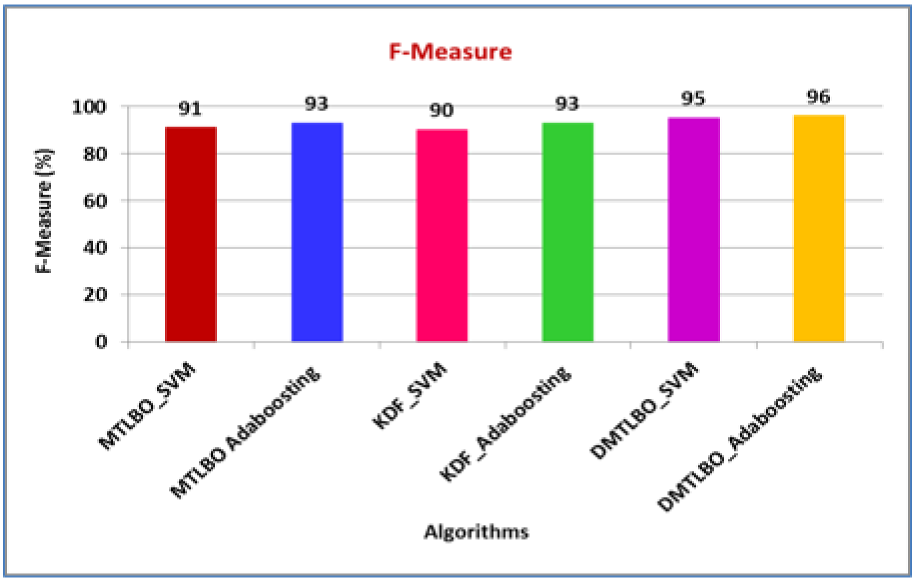

Figure 4: F-Measure

Figure 4 presents F-Measure of the standard schemes. The propounded DMTLBO_Adaboosting mechanism yields 5%, 3%, 6%, 2% and 1% improved F-Measure in contrast to standard schemes. Figure 5 shows the Time Period of the benchmarked schemes. The propounded DMTLBO_Adaboosting scheme offers 65%, 39%, 87%, 61% and 26% better Time Period when compared to the standard schemes. Figure6 shows the Error Rate of the standard schemes. The propounded DMTLBO_Adaboosting scheme involves 75%, 58%, 83%, 75% and 33% reduced Error Rate when compared to the benchmarked schemes.

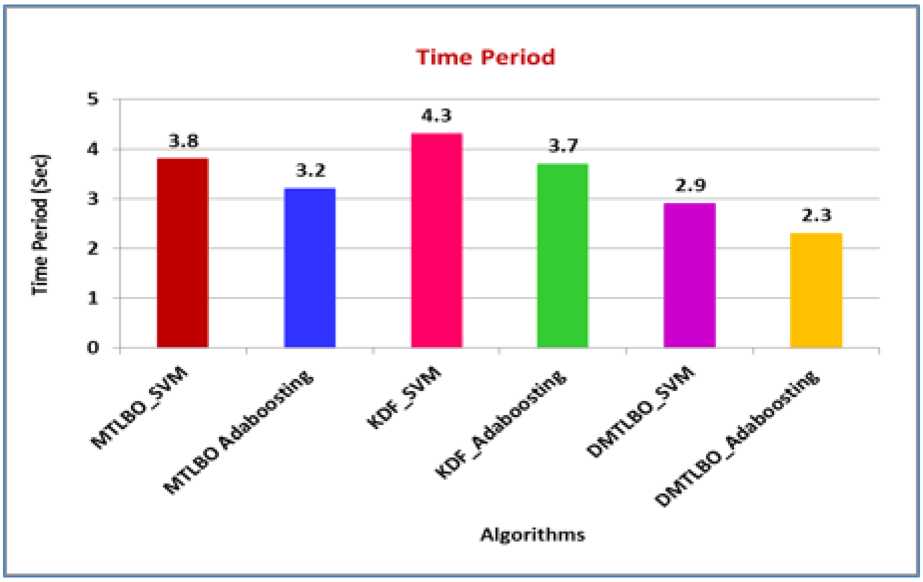

Figure 5: Time Period

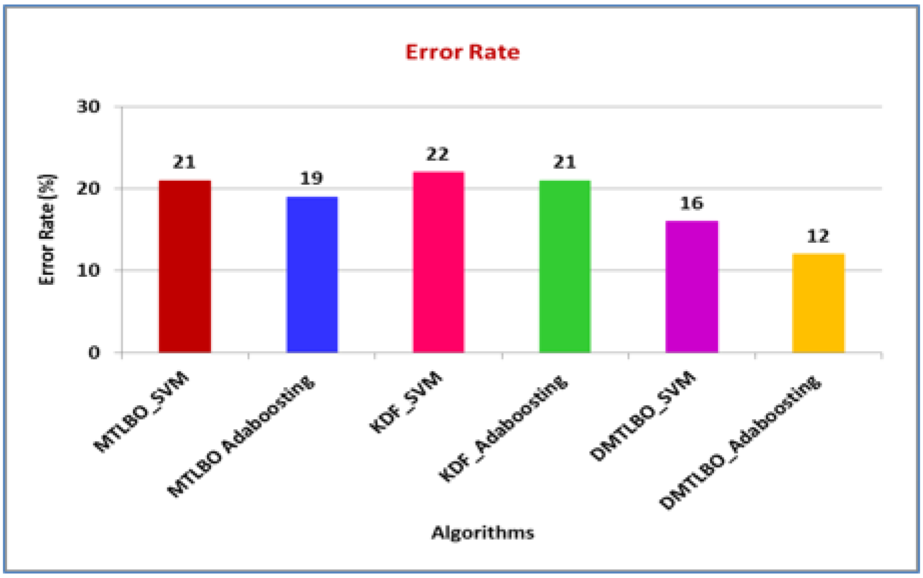

Figure 6: Error Rate

5.2 Wisconsin Diagnostic Breast Cancer (WDBC)

Wisconsin Diagnostic Breast Cancer (WDBC) is a dataset for Breast cancer analysis. It includes around 699 instances, wherein 458 are benign, while 241 are malignant. 11 attributes are considered.

Table 4: Number of Features for WDBC Dataset

No. of Attributes	TLO	KD	DMTLO
32	20	23	26

Table 4 shows the number of features using TLBO, KDF and DMTLBO for WDBC dataset. Table 5 shows the features selected using TLBO, KDF and DMTLBO for WDBC dataset. Table 6 shows the performance for WDBC dataset.

Table 5: Attributes of WDBC Dataset

TLO	KD	DMTLO
12,13,11,27,28,8,7,29,6,18,17,16,19,10,15,14,22,21,26	11,12,13,14,17,27,28,29,30,15,16,17,23,22,21,18,19,2,3,4,5,24,8	12,13,11,27,28,29,26,8,7,25,9,5,18,30,17,16,19,10,15,2,14,1,22,21,6,4

Table 6: Performance of WDBC Dataset

Algorithm	Accuracy	Precision	Recall	F-Measure	Time Period	Error Rate
MTLBO_SVM	93	91	92	93	4.7	22

MTLBO Adaboosting	97	94	95	95	4.2	20
KDF_SVM	93	92	92	90	4.3	18
KDF_Adaboosting	95	93	93	95	3.9	14
DMTLBO_SVM	96	95	96	95	3.2	16
DMTLBO_Adaboosting	98	96	97	96	2.7	12

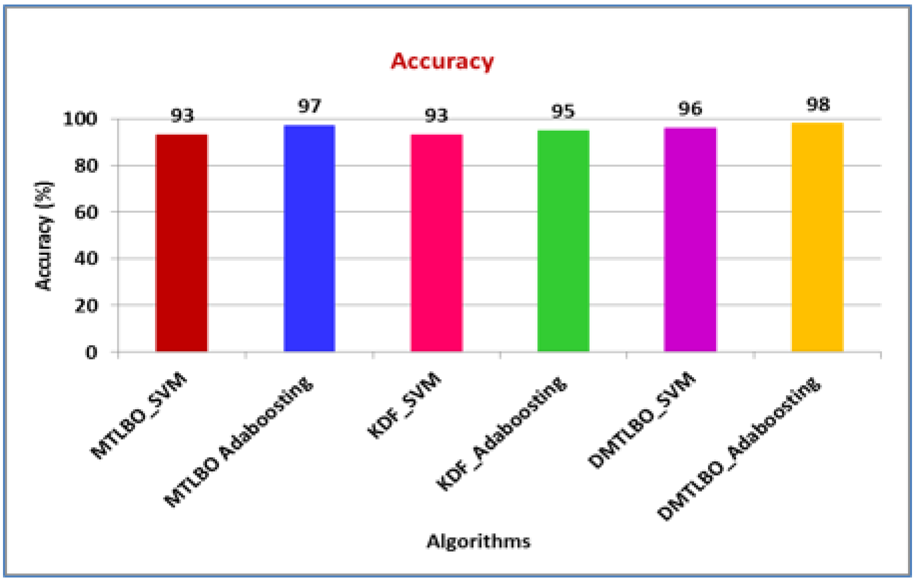

Figure 7: Accuracy

Figure 7 shows the Accuracy of MTLBO_SVM, MTLBO benchmarked schemes. The propounded DMTLBO_Adaboosting scheme offers 5%, 1%, 5%, 3% and 2% improved Accuracy when compared to the standard schemes. Figure 8 shows the Precision of the standard schemes. The propounded DMTLBO_Adaboosting scheme offers 5%, 2%, 4%, 3% and 1% enhanced Precision when compared to the benchmarked schemes.

Figure 9 shows the Recall of the benchmarked schemes. The propounded DMTLBO_Adaboosting scheme offers 5%, 2%, 5%, 4% and 1% improved Recall when compared to the standard schemes. Figure 10 shows the F-Measure of the benchmarked schemes. The propounded DMTLBO_Adaboosting scheme offers 3%, 1%, 6%, 2% and 1% better F-Measure when compared to the standard schemes.

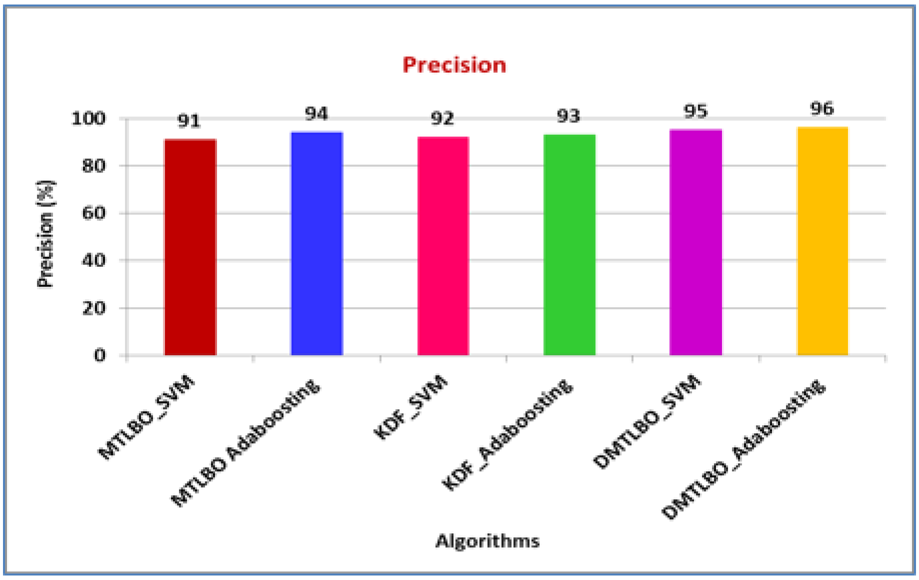

Figure 8: Precision

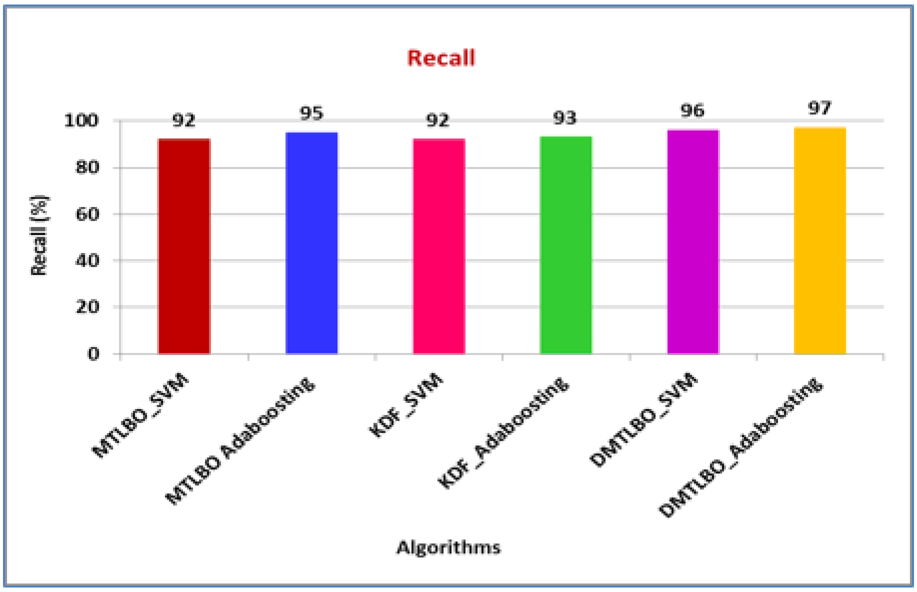

Figure 9: Recall

Figure 11 shows the Time Period of the standard schemes. The propounded DMTLBO_Adaboosting scheme offers 81%, 62%, 63%, 33% and 24% better Time Period when compared to the benchmarked schemes. Figure 12 presents the Error rate of the benchmarked schemes. The propounded DMTLBO_Adaboosting scheme involves 83%, 67%, 50%, 17% and 33% reduced Error Rate when compared to the standard schemes.

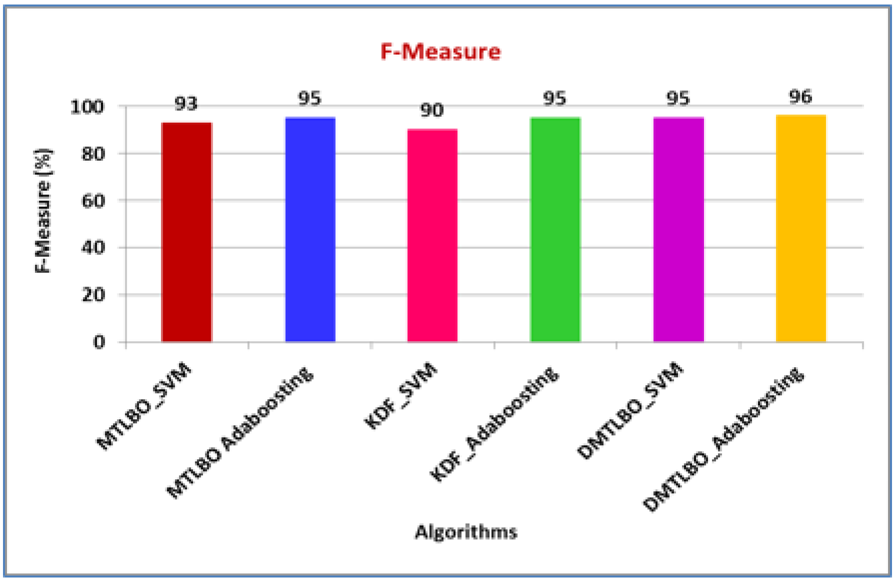

Figure 10: F-Measure

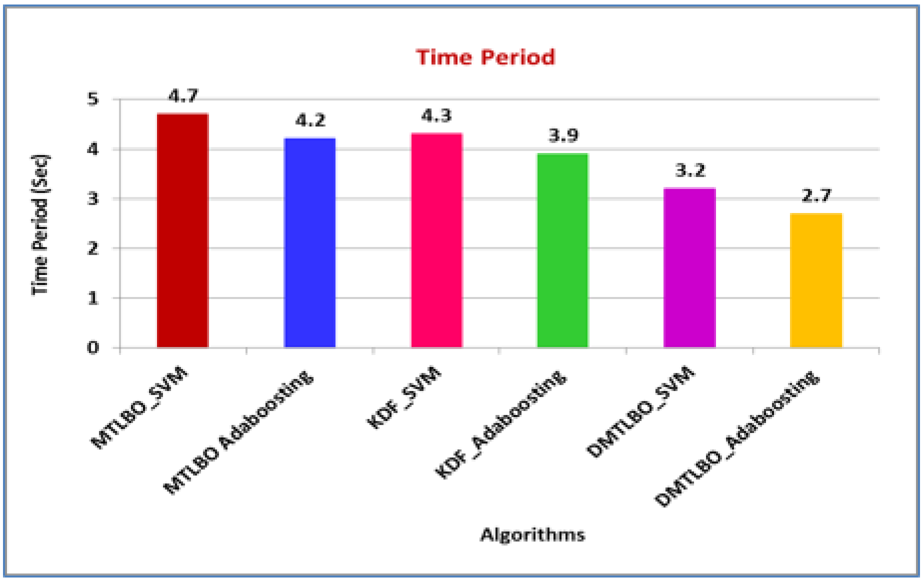

Figure 11: Time Period

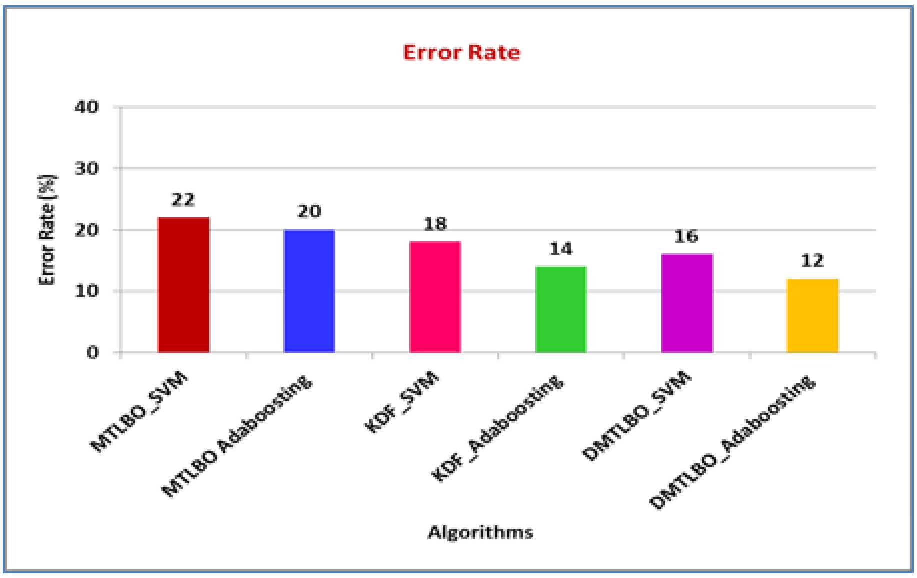

Figure 12: Error Rate

6. Conclusion

In this Thesis, TLBO, MTLBO, KDF and DMTLBO are used for selecting features in CKD dataset and WDBC. SVM and EL are used for classification.

For the CKD dataset, the propounded DMTLBO_Adaboosting scheme offers 6%, 3%, 7%, 4% and 2% improved Accuracy when compared to MTLBO_SVM, MTLBO Adaboosting, KDF_SVM, KDF_Adaboosting, DMTLBO_SVM and DMTLBO_Adaboosting schemes. Similarly, the propounded scheme offers 6%, 3%, 7%, 5% and 2% improved Precision when compared to the standard schemes. The DMTLBO_Adaboosting scheme offers 5%, 3%, 7%, 4% and 2% improved Recall when compared to the standard schemes. The propounded scheme yields 5%, 3%, 6%, 2% and 1% better F-Measure in contrast to benchmarked mechanisms. The DMTLBO_Adaboosting scheme offers 65.2%, 39.1%, 87%, 60.9% and 26.1% better Time Period when compared to the standard schemes. The propounded scheme involves 75%, 58.3%, 83.3%, 75% and 33.3% lesser Error Rate when compared to the benchmarked schemes.

Similarly, for the WDBC dataset, the propounded DMTLBO_Adaboosting scheme offers 5%, 2%, 4%, 3% and 1% improved Precision when compared to the benchmarked schemes. The DMTLBO_Adaboosting scheme offers 5%, 2%, 5%, 4% and 1% better Recall when compared to the standard schemes. The propounded scheme offers 3%, 1%, 6%, 2% and 1% better F-Measure when compared to the standard schemes. The DMTLBO_Adaboosting scheme involves 81%, 62%, 63%, 33% and 24% lesser Time Period when compared to the benchmarked schemes. The propounded scheme involves 83%, 67%, 50%, 17% and 33% lesser Error Rate in contrast to the standard schemes.In the future, this scheme can be applied to different disease datasets and the performance can be analysed.

References

1. Saxena, K., & Sharma, R. (2016). Efficient heart disease prediction system. Procedia Computer Science, 85, 962-969.
2. Hsiao, H. C., Chen, S. H., & Tsai, J. J. (2016, October). Deep learning for risk analysis of specific cardiovascular diseases using environmental data and outpatient records. In 2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE) (pp. 369-372). IEEE
3. Dangi, G., Choudhury, T., & Kumar, P. (2016, December). A smart approach to diagnose heart disease through machine learning and springleaf marketing response. In 2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE) (pp. 1-6). IEEE.
4. Sharma, H., & Rizvi, M. A. (2017). Prediction of heart disease using machine learning algorithms: A survey. International Journal on Recent and Innovation Trends in Computing and Communication, 5(8), 99-104.
5. Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017, July). A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In 2017 IEEE Symposium on Computers and Communications (ISCC) (pp. 204-207). IEEE.
6. Karthick, D., & Priyadharshini, B. (2018, January). Predicting the chances of occurrence of Cardio Vascular Disease (CVD) in people using classification techniques within fifty years of age. In 2018 2nd International Conference on Inventive Systems and Control (ICISC) (pp. 1182-1186). IEEE.
7. Mathan, K., Kumar, P. M., Panchatcharam, P., Manogaran, G., & Varadharajan, R. (2018). A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease. Design automation for embedded systems, 22(3), 225-242.
8. Rathnayakc, B. S. S., & Ganegoda, G. U. (2018, April). Heart diseases prediction with data mining and neural network techniques. In 2018 3rd International Conference for Convergence in Technology (I2CT) (pp. 1-6). IEEE.
9. Peili, Y., Xuezhen, Y., Jian, Y., Lingfeng, Y., Hui, Z., & Jimin, L. (2018, April). Deep learning model management for coronary heart disease early warning research. In 2018 IEEE 3rd

International Conference on Cloud Computing and Big Data Analysis (ICCCBDA) (pp. 552-557). IEEE.

10. Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In 2020 international conference on electrical and electronics engineering (ICE3) (pp. 452-457). IEEE.
11. Princy, R. J. P., Parthasarathy, S., Jose, P. S. H., Lakshminarayanan, A. R., & Jeganathan, S. (2020, May). Prediction of Cardiac Disease using Supervised Machine Learning Algorithms. In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 570-575). IEEE.
12. Chen, J. I. Z., & Hengjinda, P. (2021). Early Prediction of Coronary Artery Disease (CAD) by Machine Learning Method-A Comparative Study. Journal of Artificial Intelligence, 3(01), 17-33.
13. El-Bialy, R., Salamay, M. A., Karam, O. H., & Khalifa, M. E. (2015). Feature analysis of coronary artery heart disease data sets. Procedia Computer Science, 65, 459-468.
14. Long, N. C., Meesad, P., & Unger, H. (2015). A highly accurate firefly based algorithm for heart disease prediction. Expert Systems with Applications, 42(21), 8221-8231.
15. Chavan Patil, A. B., & Sonawane, P. (2017). To Predict Heart Disease Risk and Medications Using Data Mining Techniques with an IoT Based Monitoring System for Post-Operative Heart Disease Patients. International Journal on Emerging Trends in Technology (IJETT), 4, 8274-8281.
16. Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, 36, 82-93.
17. Aggrawal, R., & Pal, S. (2020). Sequential feature selection and machine learning algorithm-based patient's death events prediction and diagnosis in heart disease. SN Computer Science, 1(6), 1-16
18. Javeed, A., Rizvi, S. S., Zhou, S., Riaz, R., Khan, S. U., & Kwon, S. J. (2020). Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification. Mobile Information Systems, 2020.
19. Shah, S. M. S., Shah, F. A., Hussain, S. A., & Batool, S. (2020). Support vector machines-based heart disease diagnosis using feature subset, wrapping selection and extraction methods. Computers & Electrical Engineering, 84, 106628.
20. Valarmathi, R., & Sheela, T. (2021). Heart disease prediction using hyper parameter optimization (HPO) tuning. Biomedical Signal Processing and Control, 70, 103033.
21. Magesh, G., & Swarnalatha, P. (2021). Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction. Evolutionary Intelligence, 14(2), 583-593.

22. Bhuvaneswari, R., & Kalaiselvi, K. (2012). Naive Bayesian classification approach in healthcare applications. International Journal of Computer Science and Telecommunications, 3(1), 106-112.
23. Waghulde, N. P., & Patil, N. P. (2014). Genetic neural approach for heart disease prediction. International Journal of Advanced Computer Research, 4(3), 778.
24. Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017, July). A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In 2017 IEEE Symposium on Computers and Communications (ISCC) (pp. 204-207). IEEE.
25. Wang, H., Shi, H., Chen, X., Zhao, L., Huang, Y., & Liu, C. (2020). An improved convolutional neural network based approach for automated heartbeat classification. Journal of medical systems, 44(2), 1-9.