Webology (ISSN: 1735-188X)
Volume 18, Number 3, 2021

Csiszar-Morimoto Sammon Projective Feature Selection Based Modest Adaboost Classification For Heart Disease Diagnosis With Big Data

Dr. D. Nalini

Assistant Professor, Department of Computer Science, Kurinji College of Arts and Science (Affiliated to Bharathidasan University), Trichy-2

ABSTRACT

Big data is a group of large data that grow exponentially with time. Heart disease is the most risky and life snatching chronic disease all over the world. The heart not supplies sufficient blood to additional body parts. Many prediction techniques were introduced for finding the heart disease with better performance. But, the existing techniques failed to improve the accuracy and reduce the time consumption. In order to address these problems, Csiszar-Morimoto Sammon Projective Feature Selection based Modest AdaBoost Classifier (CMSPFS-MABC) Model is introduced. The main objective of the proposed CMSPFS-MABC Model is to perform an efficient heart disease diagnosis with minimal time consumption and higher accuracy. CMSPFS-MABC Model performed csiszar-morimoto divergence feature selection and pearson correlated hoeffding steep descent modest adaboost classification. Csiszar-Morimoto divergence is determined between the objective functions and feature. When divergence is higher than threshold, then Sammon function projects the features into similar subset. Otherwise, feature is projected into dissimilar subset. After that, the base classifier considers the training patient data and feature information. The patient data are trained with Hoeffding decision tree classifier to identify the correlation between training patient data and testing patient data. Depending on correlation measure, abnormal patient data and normal patient data are diagnosed. The base classifiers are combined using modest adaboost ensemble classifier. The similar weight is assigned to attain the base classification results. The error is determined for every base classifier. The weight is informed depending on the computed error value. Lastly, the classifier with lesser error is selected as best classifier in CMSPFS-MABC model for improving the disease diagnosing accuracy. Experimental evaluation is conducted for performance analysis of the proposed CMSPFS-MABC model using the Cardiovascular Disease dataset with different metrics like disease diagnosing accuracy, disease diagnosing time and error rate. The discussed result shows that the proposed technique improves the accuracy of heart disease diagnosis and reduces the time consumption as well as error rate than the conventional prediction methods.

Keywords: heart disease, Hoeffding decision tree classifier, Sammon function, Csiszar-Morimoto divergence, modest adaboost ensemble classifier, prediction

1. INTRODUCTION

Big data is the large area used to examine and extract the information. Heart disease is the significant human disease in world that affected the human life. In heart disease, heart failed to push required amount of blood to additional parts of body. Many machine learning algorithms were introduced to address the heart disease diagnosis problems. Neural Network model was introduced in [1] to predict the heart disease class with efficient features. The designed model minimized the error rate and increased the accuracy. But, the time consumption was not minimized by designed model. An automatic diagnostic method was designed in [2] for clinical heart disease diagnosis. The designed method identified the relevant feature subset through feature selection and extraction methods. Radial basis function kernel-based support vector machine was introduced to categorize the human as the heart disease patient (HDP) or normal patient. But, the error rate was not minimized by designed method [13].

A big health application system was introduced in [3] depending on optimal artificial neural network (OANN) for heart disease diagnosis. OANN comprised distance based misclassified instance removal (DBMIR) and learning based optimization (TLBO) algorithm. However, the detection accuracy was not improved by OANN. A predictive method was introduced in [4] for heart disease diagnosis with help of machine learning methods. Principal Component Analysis (PCA) and Fuzzy Support Vector Machine (FSVM) were introduced for missing value computation. PCA and FSVM were used to minimize the computation time for disease prediction. Though the time consumption was reduced, the computational complexity was not minimized.

A multi-task deep and wide neural network (MT-DWNN) was introduced in [5] for forecasting the fatal complication during hospitalization. MT-DWNN model attained better prediction performance on renal dysfunction in HF patients. But, the error rate was not minimized by MT-DWNN model. Radial Basis Function (RBF) neural network classification algorithm was introduced in [6] depending on analysis and nearest neighbor propagation (AP) algorithm. The similarity matrix was varied through exponential function to improve the classification accuracy and convergence speed. Though the classification accuracy was improved, the time complexity was not improved by RBF neural network classification algorithm.

An imperialist competitive algorithm with meta-heuristic approach was introduced in [7] to choose the prominent heart disease features. The designed algorithm was employed to provide the optimal response for feature selection. But, the error rate was not minimized by designed algorithm. A beetle swarm optimization and adaptive neuro-fuzzy inference system (BSO-ANFIS) model was introduced in [8] for heart disease and multi-disease diagnosis. The modified crow search algorithm was employed for feature extraction and ANFIS classification model were optimized by BSO algorithm. However, the computational complexity was not minimized by BSO-ANFIS model.

A pathological factors for early HF detection was carried out in [9] for social network based approach. The electronic health records were employed to determine the risk factor similarity. The similarity values were determined to construct the unweighted and weighted medical social network. But, the detection time was not minimized. A new deep neural network method was introduced in [10] for heart disease detection. The feature vector was employed to discriminate the cardiac sound through gathering set of feature from cardiac sounds. But, the feature selection was not carried out for heart disease detection.

A convolution neural network model was designed in [11] to identify and categorize cardiac arrhythmias with ECG dataset given by China Physiological Signal Challenge (CPSC). However, the time consumption was not minimized by designed model. Non-linear Iterative Partial Least Squares was designed in [12] to perform data dimensionality reduction. Self-Organizing Map technique was employed for task clustering and Neuro-Fuzzy Inference System was employed for hepatitis disease prediction. But, the computational cost was not minimized by Non-linear Iterative Partial Least Square.

The above existing techniques are reviewed and identified the issues such as higher heart disease diagnosing accuracy, lesser heart disease diagnosing time, higher computational cost, and lesser feature selection performance [14]. In order to address these problems, the proposed CMSPFS-MABC Model is introduced in this article.

The main contribution of the article is given as: Csiszar-Morimoto Sammon Projective Feature Selection based Modest AdaBoost Classifier (CMSPFS-MABC) Model is introduced for efficient heart disease diagnosis with minimal time consumption. CMSPFS-MABC Model performed csiszar-morimoto divergence feature selection and pearson correlated hoeffding steep descent modest adaboost classification. Csiszar-Morimoto divergence is determined between objective function and features. Sammon function classified the features into similar subset and dissimilar subset. The patient data are trained through Hoeffding decision tree classifier to recognize the correlation between training and testing patient data. With correlation measure, abnormal patient data and normal patient data are classified. The base classifiers are joined with modest adaboost ensemble classifier to reduce the error rate.

The remaining part of the article is structured as follows. In Section 2, the architecture diagram of CMSPFS-MABC Model is explained with neat diagram. It addition, it also discusses csiszar-morimoto diverged sammon projection and pearson correlated hoeffding modest adaboost classification. In Section 3, experimental results are discussed with performance metrics in detail. Section 4 discussed the result of existing and proposed methods. The final Section 5 is concerned with conclusion of the paper.

2. METHODOLOGY

Heart disease is the main cause of death for both men and women. Heart disease denotes group of conditions that affect the heart muscle or nearby arteries which supply heart with blood. The heart disease identification is carried out through classification methods. The existing classification algorithm categorized the patient data to identify the heart disease. However, the disease diagnosing accuracy performance was not improved by existing methods. In order to increase the disease diagnosing accuracy and to reduce the prediction time, an effective ensemble classifier termed Csiszar-Morimoto Sammon Projective Feature Selection based Modest AdaBoost Classifier (CMSPFS-MABC) Model is introduced. In the proposed CMSPFS-MABC Model, feature selection and classification process is carried out using machine learning techniques. The feature selection is carried out to select the relevant features for improving the heart disease diagnosing performance. After that, modest adaboost ensemble classifier is used for performing the accurate heart disease prediction. Hoeffding decision trees are considered as the weak classifier for diagnosing the heart disease with the training data samples. Modest AdaBoost is a boosting technique that joins all weak learners and provides the strong classification results with lesser error. The architecture diagram of the proposed technique is introduced in figure 1.

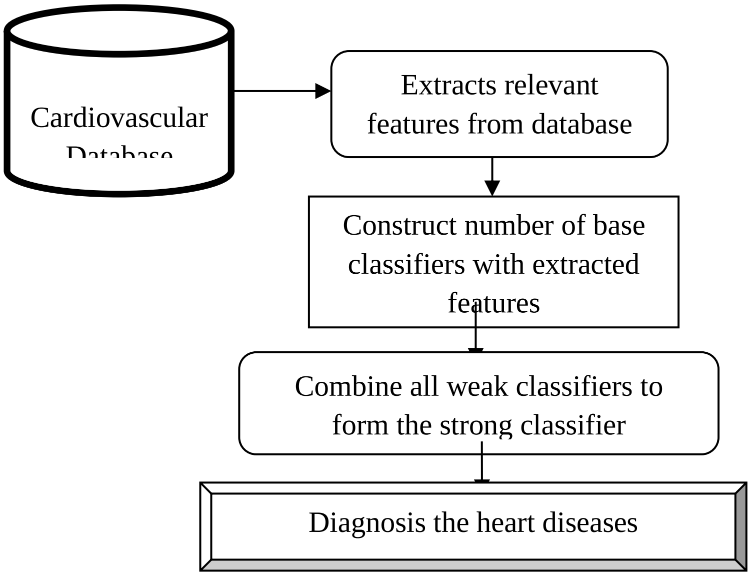
Cardiovascular
Database
Extracts relevant features from database
Construct number of base classifiers with extracted features
Combine all weak classifiers to form the strong classifier
Diagnosis the heart diseases

Figure 1 Architecture diagram of CMSPFS-MABC Model

Figure 1 illustrates the architecture diagram of CMSPFS-MABC Model for diagnosing the heart disease by using ensemble classifier. The number of patient data and their feature information is collected from the cardiovascular dataset. Depending on the selected feature information, the normal and abnormal patient data are classified to identify the heart disease at the starting stage. The following section explains the feature selection and ensemble classifier in brief manner.

2.1 Csiszar-Morimoto Diverged Sammon Projection

The proposed CMSPFS-MABC model performs the relevant feature selection for accurate disease prediction with lesser time consumption. Feature selection is an essential process employed to identify the relevant and necessary features from input dataset. Sammon projection is a machine learning algorithm used to project the high-dimensional space to lower dimensionality space through maintaining the inter-point distance structure. The similar features are identified through neighboring points and dissimilar features are identified through the distant points with higher probability. The similar and dissimilar features are recognized with help of Csiszar-Morimoto divergence. The Csiszar-Morimoto divergence is defined as the distance between two points. The structural diagram of feature selection is shown in figure 2.

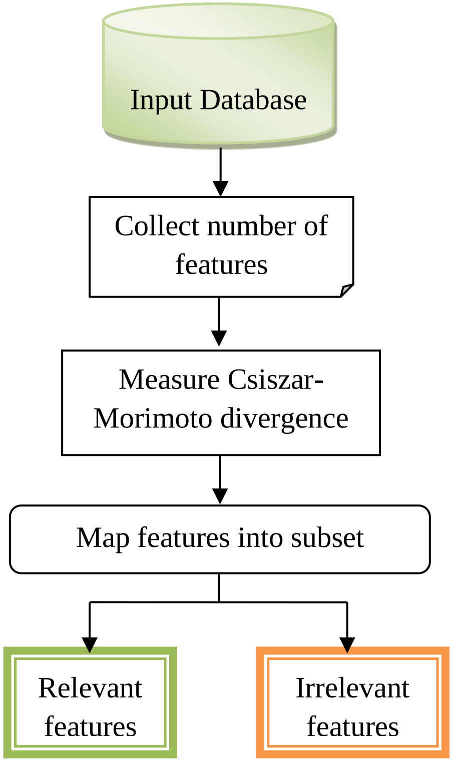

Figure 2 Structural Diagram of Feature Selection Process

Figure 2 explains the structural diagram of relevant feature selection process. Let us take, the number of features '$\mathrm{Fe_i} = f_1, f_2, .., f_n$' from given input dataset in the high dimensional space. Through using Csiszar-Morimoto divergence, the distance is computed between the input feature from dataset and objective. Csiszar-Morimoto divergence is computed as,

$$\varphi_{CM} = \left\|\mathrm{Fe_i} - O_j\right\| \qquad (1)$$

From (1), 'φ_{CM}' represents the Csiszar-Morimoto divergence, '$\mathrm{Fe_i}$' symbolizes the features from input dataset, 'O_j' denotes the objective. The divergence value ranges between 0 and 1. After that, the threshold value is predefined to map the input features into different subsets. Consequently, the projection output is given as,

$$\mathrm{P} \rightarrow \begin{cases} \varphi_{\mathrm{CM}} > \gamma \,; \text{Irrelevant features} \\ \varphi_{\mathrm{CM}} < \gamma \,; \text{Relevant features} \end{cases} \quad (2)$$

From (2), ‘P’ represents the projection function, ‘γ’ denotes the threshold. When the divergence of the feature is higher than threshold value, feature is considered as the dissimilar feature. Otherwise, the feature is considered as the relevant feature. By this manner, the similar features are mapped into low-dimensional space. The algorithmic process of Csiszar-Morimoto Diverged Sammon Projection is given as,

```
Input: Dataset ‘$Dt$’, Number of features $Fe_i = f_1, f_2, .., f_n$
Output: Improves feature selection performance
Begin
    Number of features $Fe_i = f_1, f_2, .., f_n$ taken as input at input layer
    For each feature $Fe_i \in Dt$
        Measure the Csiszar-Morimoto divergence between features ‘$Fe_i$’ and objective ‘$O_j$’
        if ( $\varphi_{CM} > \gamma$ ) then
            Project the features as similar
        else
            Project the features as dissimilar
    End if
    Select the similar feature subset
    Remove the dissimilar feature subset
    End for
End
```

Algorithm 1 Csiszar-Morimoto Diverged Sammon Projection Algorithm

Algorithm 1 explains the step-by-step process of feature selection with higher accuracy. The features are considered as input. Csiszar-Morimoto divergence is determined between the objective function and feature in CMSPFS-MABC model. When the divergence is higher than threshold, then Sammon function projects the features into similar subset. Otherwise, the feature is projected into dissimilar subset. By this way, efficient feature selection is carried out with minimal time consumption. The brief discussion about the Pearson Correlated Hoeffding Modest AdaBoost Classification is given in next sub-section.

2.2 Pearson Correlated Hoeffding Modest AdaBoost Classification

The classification is the process categorizing the patient data into two different classes, namely normal class and abnormal class. The Pearson Correlated Hoeffding Steep Descent Modest AdaBoost Classifier in CMSPFS-MABC model performs the classification by constructing the strong classifier. Pearson Correlated Hoeffding Steep Descent Modest AdaBoost Classifier constructs the number of weak classifiers as a hoeffding decision trees to categorize the patient data. The base classifier results are joined to obtain the strong classification results using modest adaboost classifier. Initially, the relevant features from number of patient data in dataset is expressed as follows,

$$P_{D_k} \rightarrow f_1, f_2, f_3, \dots f_n \in Dt \qquad (3)$$

From (3), $f_1, f_2, f_3, \dots f_n$ denotes features in the dataset 'Dt'. With given input patient data and their feature information, the hoeffding decision trees are constructed for classification. In proposed technique, hoeffding decision tree constructs the classification models in form of tree structure. It partitions the dataset into two smaller subsets. The tree comprises the decision nodes and leaf nodes. Decision node comprises the two or more branches. Leaf node denotes the class labels. The uppermost decision node in the hoeffding decision tree is the root node. The key benefit of the hoeffding decision tree classifier is to manage the categorical and numerical patient data.

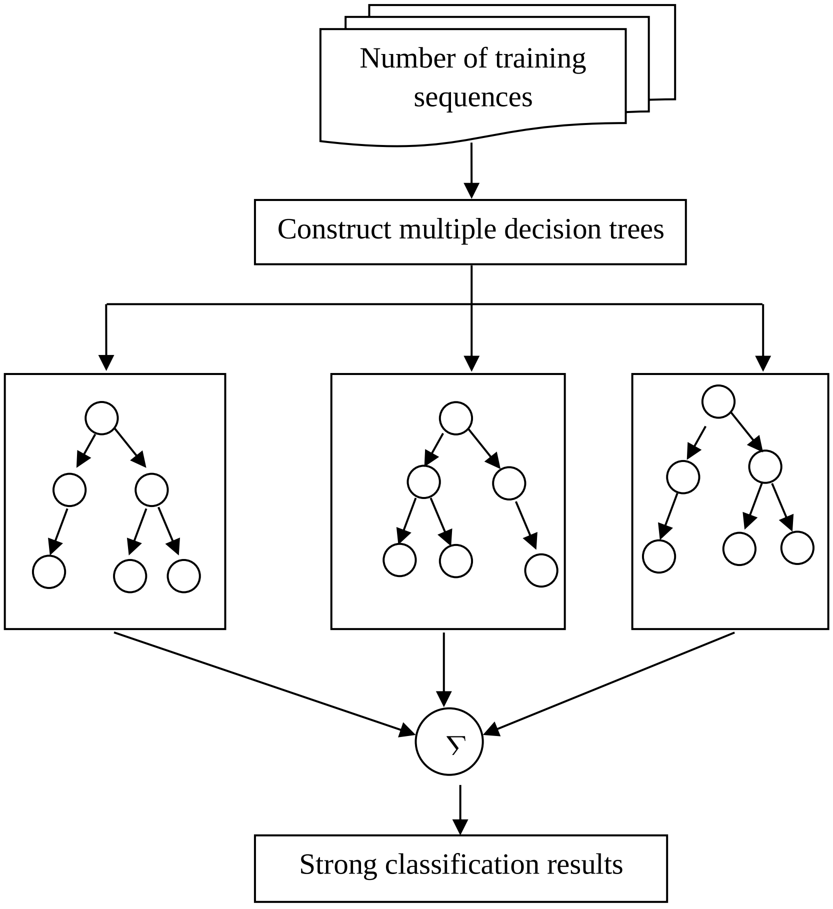
Number of training sequences
Construct multiple decision trees
Σ
Strong classification results

Figure 3 Process of Pearson Correlated Hoeffding Steep Descent Modest AdaBoost Classification

Figure 3 explains the flow process of Pearson Correlated Hoeffding Steep Descent Modest AdaBoost Classification. The ensemble classification result provides accurate heart disease diagnosis with minimum error rate. Initially, the number of weak classifiers (i.e., weak classifiers) is constructed and the patient data is classified into normal or abnormal. The decision tree classifier performs the correlation analysis between the training sequences and testing sequences with the number of features in the dataset to categorize the training sequences. The relationship between two sequences is determined by pearson product-moment correlation coefficient. Pearson product-moment correlation is formulated as,

$$r_{pd_{tr},pd_{ts,=}} \frac{m*\sum pd_{tr}*pd_{ts}-(\sum pd_{tr})(\sum pd_{ts})}{\sqrt{[m*\sum {pd_{tr}}^2-(\sum pd_{tr})^2][m*\sum {pd_{ts}}^2-(\sum pd_{ts})^2]}} \qquad (4)$$

From (4), 'r' represent the correlation coefficient. 'pd_{tr}' denotes the training patient data. 'pd_{ts}' symbolize the testing patient data. 'm' denotes the number of patient data. $\sum pd_{tr} * pd_{ts}$ denote the sum of the product of paired score of two patient data. ${pd_{tr}}^2$ symbolize the squared score of 'pd_{tr}'. '${pd_{ts}}^2$' denotes squared score of 'ps_{tr}'. After determining the correlation measure, coefficient provides two results like "+1" and "-1". When the coefficient provide the positive correlation between the training and testing patient data, the output result obtained is '+1'. Otherwise, the result obtained is '-1'. Therefore, the base classifier categorizes the patient data as the abnormal patient data. It is given by,

$$r_{pd_{tr},pd_{ts}} = \begin{cases} +1, & normal\ patient\ data \\ -1, & abormal\ patient\ data \end{cases} \qquad (5)$$

Based on the classification results, accurate heart disease diagnosis is obtained using base classifiers. In addition, Pearson Correlated Hoeffding Modest AdaBoost Classifier Classification is used in CMSPFS-MABC model to improve the base classifier performance. The boosting classifier joins the base learners and attains the strong classification results through reducing the loss function (i.e. error). At last, strong learner efficiently categories the patient data as normal or abnormal and reduces the incorrect classification.

Let us take, number of base learner results as '$\mathrm{H}(1), \mathrm{H}(2), .. \mathrm{H}(\mathrm{l})$'. The Pearson Correlated Hoeffding Steep Descent Modest AdaBoost Classification in CMSPFS-MABC model initializes the similar weight 'ω' to every base classifier. It is given by,

$$\omega \rightarrow \{\mathrm{H}(1), \mathrm{H}(2), .. \mathrm{H}(\mathrm{l})\} \qquad (6)$$

The strong classifier joins all the base learners with their weight value. It is formulated as,

$$\mathrm{SC} = \sum_{\mathrm{u}=1}^{\mathrm{l}} \omega * \mathrm{H}_{\mathrm{u}} \qquad (7)$$

From (7), ‘SC’ represents the strong classifier, ‘ω’ symbolizes the weight of every base classifier ‘H_u’. After that, the error is computed for every base learner and given by,

$$\text{Error} = \text{observed value} - \text{actual value} \qquad (8)$$

From (8), ‘Error’ is determined. After calculating error value, the initial weight of base classifier is revised. The updated weight is given by,

$$\omega' \rightarrow \{H(1), H(2), .. H(l)\} \qquad (9)$$

From (9), ‘ω'’ represents the updated weight of every classifier based on error. The weight of the every base classifier is enhanced when the patient data are incorrectly classified. The weight is reduced when the base leaner correctly classified the patient data. Accordingly, the weight is allocated with the error value. Then, the ensemble classifier employs the steepest descent function to identify the best classifier with lesser error and given as,

$$\text{steepest descent function} = \arg\min \mathrm{E\ rror}\,(H_u) \qquad (10)$$

From equation (10), modest adaboost classifier chooses the base learner H_u with lesser weight value as strong classification results. It is formulated as,

$$Y = \sum_{u=1}^{l} \omega' \, H_u \qquad (11)$$

From (11), ‘Y’ symbolizes the final classification result of modest adaboosting ensemble classifier with minimum updated weight ‘ω'’. Accordingly, the strong classifier efficiently detects the normal patient data and abnormal patient data. In addition, the steepest descent is employed to identify the best classifier with lesser error for attaining higher accuracy and lesser error rate. Based on classification results, the accurate heart disease diagnosis is performed with minimum time. The algorithmic process of Pearson Correlated Hoeffding Steep Descent Modest AdaBoost Classification is given as follows,

```
Input: Input dataset ‘Dt’, features ‘$f_1, f_2, f_3, \dots f_n$’
Output: Improve heart disease diagnosis performance
Begin
    Construct hoeffding decision tree with training sequences
    Perform correlation based classification using $r_{pd_{tr},pd_{ts}}$
    If ( $r_{pd_{tr},pd_{ts}} = +1$) then
        Patient data is classified as normal
    else ( $r_{ps_{tr},ps_{ts}} = -1$) then
        Patient data are classified as abnormal
    end if
    Apply modest adaboosting classifier and combines all base classifier
    Initialize similar weights ω to $H_u$
```

```
For each $H_u$
    Calculate the Error
    Update initial weight weight $\omega'$ of $H_u$
    Choose the base learner with minimum weight as strong learner
    Obtain strong classification results
End for
End
```

Algorithm 2 Pearson Correlated Hoeffding Steep Descent Modest AdaBoost Classification

Algorithm 2 explains the ensemble classification to identify the heart disease at starting stage. The base classification and strong classification results are performed with features of the training and testing patient data. At first, the base classifier considers the input as training patient data and feature information from the cardiovascular dataset. The patient data are trained using Hoeffding decision tree classifier to identify the correlation between the training patient data and testing patient data. Depending on the correlation measure, the abnormal patient data and normal patient data are diagnosed. The base classifiers are joined using modest adaboost ensemble classifier. The similar weight is allocated to base classification results. After that, the error is determined for every base classifier. The weight is updated based on the computed error value. Finally, the classifier with lesser error is chosen as best classifier in CMSPFS-MABC model for improving the disease diagnosing accuracy.

3. EXPERIMENTAL EVALUATION

An experimental evaluation of proposed CMSPFS-MABC model, existing methods, namely Neural Network model [1] and automatic diagnostic method [2] are implemented using Java language with Cardiovascular Disease dataset. Dataset is taken from from Kaggle https://www.kaggle.com/sulianova/cardiovascular-disease-dataset for diagnosing the patient data with the cardiovascular disease. The dataset comprises three types of input features. This dataset includes 13 attributes of three types, namely objective, examination and subjective with 70 000 patients data records. The attributes are id, age, height, weight, gender, systolic blood pressure, cholesterol, smoking etc. The final target denotes the presence or absence of cardiovascular disease. In order to conduct the experiments, number of patient data ranges from 1000 to 10000.

4. RESULT AND DISCUSSION

The experimental results of the proposed CMSPFS-MABC model and existing Neural Network model [1] and automatic diagnostic method [2] are discussed in this section with different performance metrics like disease diagnosis accuracy, error rate and disease diagnosis time.

4.1 Impact on Disease Diagnosis Accuracy:

It is described as the ratio of number of patient data that are correctly diagnosed through classification process to the total number of patient data considered as input. The disease diagnosis accuracy is determined as,

$$\mathrm{DDA} = \left(\frac{\text{Number of patient data that are correctly diagnosed}}{\mathrm{m}}\right) * 100 \qquad (11)$$

From (11), 'DDA' symbolizes the disease diagnosis accuracy, 'm' denotes the number of patient data. The disease diagnosis accuracy is determined in terms of percentage (%).

Table 1 Tabulation for Disease Diagnosis Accuracy

Number of Patient Data (Number)	Disease Diagnosis Accuracy (%)		
	Proposed CMSPFS-MABC model	Neural Network model	Automatic diagnostic method
1000	90	85	82
2000	93	88	85
3000	96	87	86
4000	91	86	89
5000	96	92	90
6000	92	84	87
7000	94	90	87
8000	95	89	87
9000	94	91	88
10000	97	91	89

As described in table 1, the disease diagnosing accuracy of the proposed CMSPFS-MABC model is compared with state-of-the-art methods. The observed results denotes that proposed CMSPFS-MABC model improves the disease diagnosing accuracy upon comparison with the other two existing methods neural network model [1] and automatic diagnostic method [2] for 10000 different patient data as shown in table 1. The Cardiovascular Disease dataset is taken to conduct the experiments. When considering the number of patient data as 1000, the proposed CMSPFS-MABC model correctly diagnosis the 901 patient data whereas [1] and [2] correctly diagnosis the 850 patient data and 824 patient data correspondingly. Therefore, the disease diagnosing accuracy of the proposed CMSPFS-MABC model is 90% and the disease diagnosing accuracy of [1] and [2] is 85% and 82% correspondingly. The ten different runs are performed with various number of patient data. The disease diagnosis accuracy comparison of existing and proposed methods is illustrated in figure 4.

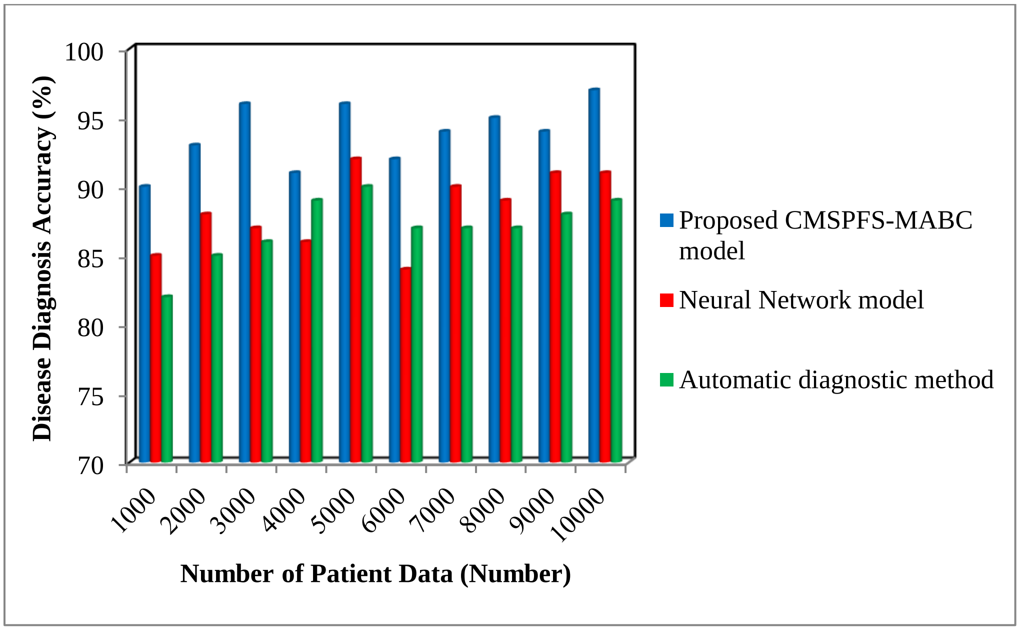

Figure 4 Measurement of Disease Diagnosing Accuracy

Figure 4 illustrates the disease diagnosis accuracy for different number of patient data. The number of patient data is described as input in horizontal axis and disease diagnosis accuracy results are shown on vertical axis. The disease diagnosis accuracy of three different methods namely CMSPFS-MABC model, existing methods neural network model [1] and automatic diagnostic method [2] is denoted by three different color cylinders like blue, red and green in the same way. The disease diagnosis accuracy results are attained that CMSPFS-MABC model outperforms better than additional two techniques. The improvement is owing to application of csiszar-morimoto divergence and pearson correlated hoeffding steep descent modest adaboost classification. Sammon function classified the features into similar subset when divergence is higher than threshold [15]. The patient data are trained through Hoeffding decision tree classifier to recognize the correlation between training and testing patient data. The base classifiers are combined using the modest adaboost ensemble classifier to improve the disease diagnosis accuracy. The average results represent that the disease diagnosis accuracy of CMSPFS-MABC model is increased by 6% and 8% when compared to existing methods.

4.2 Impact on Error rate:

It is defined as ratio of number of patient data that are incorrectly classified through classification to the total number of patient data considered as input. The error rate is determined as,

$$\mathrm{Error}_{\mathrm{Rate}} = \left(\frac{\text{Number of patient data that are incorrectly diagnosed}}{\mathrm{m}}\right) * 100 \qquad (12)$$

From (12), 'm' denotes the number of patient data. The error rate is measured in terms of percentage (%).

Table 2 Tabulation for Error Rate

Number of Patient Data (Number)	Error Rate (%)		
	Proposed CMSPFS-MABC model	Neural Network model	Automatic diagnostic method
1000	10	15	18
2000	7	12	15
3000	4	13	14
4000	9	14	11
5000	4	8	10
6000	8	16	13
7000	6	10	13
8000	5	11	13
9000	6	9	12
10000	3	9	11

As shown in table 2, the error rate of the proposed CMSPFS-MABC model is compared with existing methods. The attained results shows that proposed CMSPFS-MABC model reduces the error rate of disease diagnosis upon comparison with other two existing methods neural network model [1] and automatic diagnostic method [2] for 10000 different patient data. When considering the number of patient data as 1000, the proposed CMSPFS-MABC model incorrectly diagnosis the 99 patient data where [1] and [2] incorrectly diagnosis the 150 patient data and 176 patient data correspondingly. Consequently, the error rate of the proposed CMSPFS-MABC model is 10% and the disease diagnosing accuracy of [1] and [2] is 15% and 18% correspondingly. Ten runs are carried out with different number of patient data. The error rate comparison of existing and proposed methods is illustrated in figure 5.

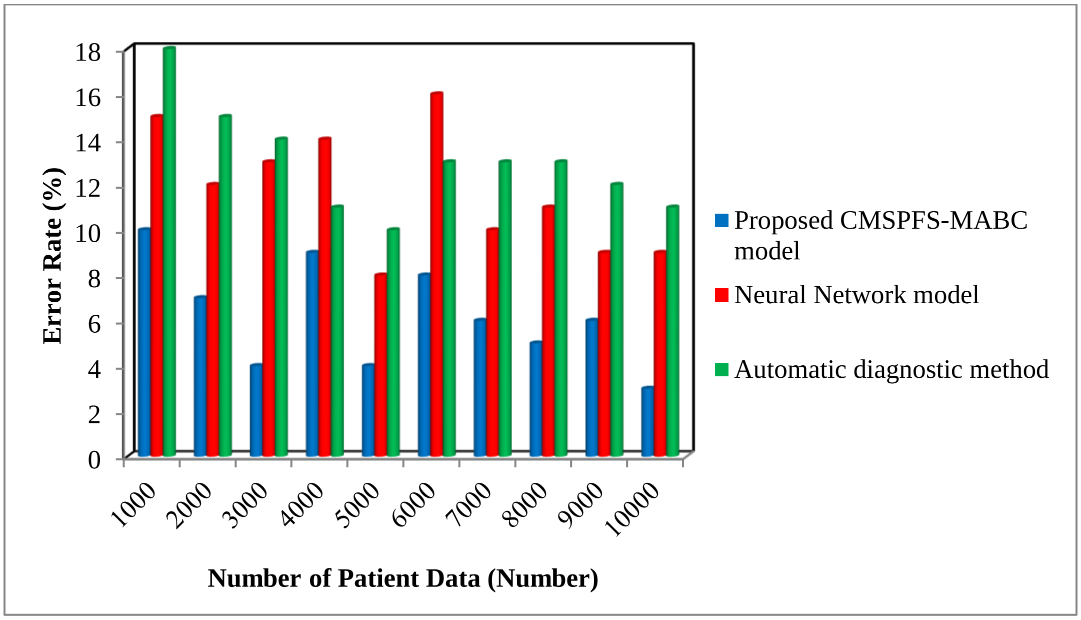

Figure 5 Measurement of Error Rate

Figure 5 shows the error rate for different number of patient data. The number of patient data is shown as input in the horizontal axis and error rate results are monitored on vertical axis. The error rate of three different methods namely CMSPFS-MABC model, existing methods neural network model [1] and automatic diagnostic method [2] is represented by three different color cylinders like blue, red and green correspondingly. The observed results identified that CMSPFS-MABC model outperforms better than additional two techniques. The enhancement is due to the application of csiszar-morimoto divergence and pearson correlated hoeffding steep descent modest adaboost classification. When divergence is larger than threshold, Sammon function classified the features into similar subset. The patient data are trained through the Hoeffding decision tree classifier to identify correlation between training and testing patient data. With help of the correlation measure, abnormal patient data and normal patient data are classified. The base classifiers are joined with help of modest adaboost ensemble classifier to reduce the error rate. The average results represent that the error rate of CMSPFS-MABC model is reduced by 47% and 52% when compared to existing methods.

4.3 Impact on Disease Diagnosis Time:

It is defined as the amount of time consumed by the algorithm for diagnosing the disease. Disease diagnosis time is the product of number of patient data and time for predicting one patient data. Therefore, the disease diagnosing time is determined as,

$$\text{DDT} = \text{m} * \text{time for predicting one patient data} \qquad (13)$$

From (13), 'DDT' represent the disease diagnosis time. 'm' denotes the number of patient data. The disease diagnosis time is measured in terms of milliseconds (ms).

Table 3 Tabulation for Disease Diagnosis Time

Number of Patient Data (Number)	Disease Diagnosis Time (ms)		
	Proposed CMSPFS-MABC model	Neural Network model	Automatic diagnostic method
1000	20	24	30
2000	23	27	34
3000	26	31	38
4000	30	35	41
5000	33	39	46
6000	37	42	50
7000	40	45	54
8000	44	49	56
9000	48	53	60
10000	51	57	64

As illustrated in table 3, the disease diagnosis time of the proposed CMSPFS-MABC model is evaluated with the existing techniques. The results shows that proposed CMSPFS-MABC model reduces the disease diagnosis time of upon comparison with other two existing methods neural network model [1] and automatic diagnostic method [2] for 10000 different patient data. When considering the number of patient data as 1000, the proposed CMSPFS-MABC model consumes 0.02ms for performing the disease diagnosis of one patient data where [1] and [2] consumes 0.024ms and 0.03ms correspondingly for performing the disease diagnosis of one patient data. Consequently, the disease diagnosis time of the proposed CMSPFS-MABC model is 20ms and the disease diagnosing time of [1] and [2] is 24ms and 30ms correspondingly. Ten runs are carried out with different number of patient data. The disease diagnosis time comparison of existing and proposed methods is shown in figure 6.

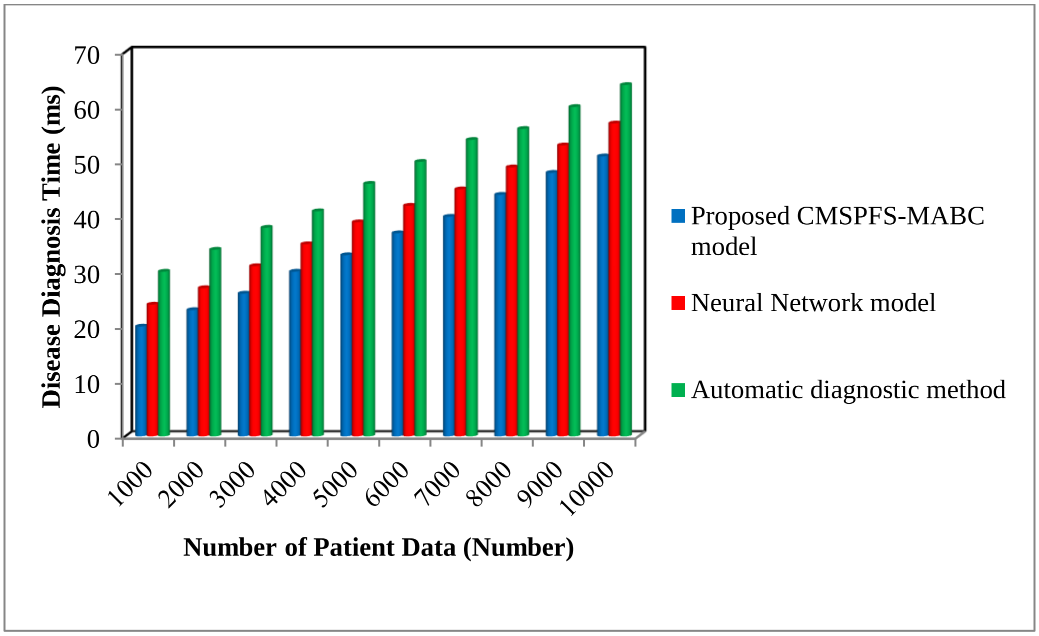

Figure 6 Measurement of Disease Diagnosis Time

Figure 6 shows the disease diagnosis time for different number of patient data. The number of patient data is given as input in horizontal axis and disease diagnosis time results are observed on vertical axis. The disease diagnosis time of three different methods namely CMSPFS-MABC model, existing methods neural network model [1] and automatic diagnostic method [2] is denoted by three different colors like blue, red and green correspondingly. The observed results noticed that the CMSPFS-MABC model outperforms well than additional two methods [19]. The improvement is due to application of csiszar-morimoto divergence and pearson correlated hoeffding steep descent modest adaboost classification. When the divergence is higher than the threshold, then Sammon function projects the features into similar subset. The patient data are trained by Hoeffding decision tree classifier to recognize the correlation between training patient data and testing patient data. Depending on the correlation measure, abnormal patient data and normal patient data are diagnosed. The base classifiers are combined using modest adaboost ensemble classifier to reduce the disease diagnosis time. The average results represent that the disease diagnosis time of CMSPFS-MABC model is reduced by 13% and 27% when compared to existing methods.

5. CONCLUSION

A new ensemble learning method called Csiszar-Morimoto Sammon Projective Feature Selection based Modest AdaBoost Classifier (CMSPFS-MABC) Model performs the heart disease diagnosis with minimal time consumption and higher accuracy [16]. Csiszar-Morimoto divergence is computed between the objective function and feature. Sammon function projects the features into similar subset. After that, patient data are trained with Hoeffding decision tree classifier to determine the correlation between the patient data and testing patient data. The base classifiers are combined through the modest adaboost ensemble classifier. This in turn helps to improve the performance the heart disease diagnosis [17]. The extensive experimental evaluation is conducted with the Cardiovascular Disease dataset. The quantitative results discussion confirms that the CMSPFS-MABC Model improves the disease diagnosis accuracy with minimal error rate and disease diagnosis time than the state-of-the-art methods[18].

REFERENCES

[1] J. Jeyaranjani, T. Dhiliphan Rajkumar and T. Ananth Kumar, "Coronary heart disease diagnosis using the efficient ANN model", Materials Today: Proceedings, Elsevier, March 2021, Pages 1-10

[2] Syed Muhammad Saqlain Shah, Faiz Ali Shah, Syed Adnan Hussain and Safeera Batool, "Support Vector Machines-based Heart Disease Diagnosis using Feature Subset, Wrapping Selection and Extraction Methods", Computers & Electrical Engineering, Elsevier, Volume 84, June 2020, Pages 1-18

[3] R. Thanga Selvi and I. Muthulakshmi, "An optimal artificial neural network based big data application for heart disease diagnosis and classification model", Journal of Ambient Intelligence and Humanized Computing, Springer, June 2020, Pages 1-11

[4] Mehrbakhsh Nilashi, Hossein Ahmadi, Azizah Abdul Manaf, Tarik A. Rashid, Sarminah Samad, Leila Shahmoradi, Nahla Aljojo and Elnaz Akbari, "Coronary Heart Disease Diagnosis through Self-Organizing Map and Fuzzy Support Vector Machine with Incremental Updates", International Journal of Fuzzy Systems, Springer, Volume 22, 2020, Pages 1376-1388

[5] Binhua Wang, Yongyi Bai, Zhenjie Yao, Jiangong Li, Wei Dong, Yanhui Tu, Wanguo Xue, Yaping Tian, Yifei Wang and Kunlun He, "A Multi-Task Neural Network Architecture for Renal Dysfunction Prediction in Heart Failure Patients with Electronic Health Records", IEEE Access, Volume 7, Pages 178392 – 178400

[6] Congshi Jiang and Yihong Li, "Health Big Data Classification Using Improved Radial Basis Function Neural Network and Nearest Neighbor Propagation Algorithm", IEEE Access, Volume 7, Pages176782-176789

[7] Jalil Nourmohammadi-Khiarak, Mohammad-Reza Feizi-Derakhshi, Khadijeh Behrouzi, Samaneh Mazaheri, Yashar Zamani-Harghalani and Rohollah Moosavi Tayebi, "New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection", Health and Technology, Springer, Volume 10, 2020, Pages 667-678

[8] Parminder Singh, Avinash Kaur, Ranbir Singh Batth, Sukhpreet Kaur and Gabriele Gianini, "Multi-disease big data analysis using beetle swarm optimization and an adaptive

neuro-fuzzy inference system", Neural Computing and Applications, Springer, 2021, Pages 1-10

[9] Chunjie Zhou, Ali Li, Aihua Hou, Zhiwang Zhang, Zhenxing Zhang, Pengfei Dai and Fusheng Wang, "Modeling methodology for early warning of chronic heart failure based on real medical big data", Expert Systems With Applications, Elsevier, Volume 151, 2020, Pages 1-13

[10] Luca Brunese, Fabio Martinelli, Francesco Mercaldo and Antonella Santone, "Deep learning for heart disease detection through cardiac sounds", Expert Systems with Applications, Elsevier, Volume 151, 2020, Pages 1-10

[11] Tsai-Min Chen, Chih-Han Huang, Edward S.C.Shih, Yu-Feng Hu, Ming-Jing Hwang, "Detection and Classification of Cardiac Arrhythmias by a Challenge-Best Deep Learning Neural Network Model", iScience, Elsevier, Volume 23, Issue 3, March 2020, Pages 1-10

[12] Mehrbakhsh Nilashi, Hossein Ahmadi, Leila Shahmoradi, Othman Ibrahim, Elnaz Akbari, "A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique", Journal of Infection and Public Health, Volume 12, Issue 1, January-February 2019, Pages 13-20.

[13] Periyasamy, R., and Nalini, D. (2015). Binary back propagation based lift association mining for heart disease and stroke identification. International journal of applied engineering research, 10(6), pp. 16071-16087.

[14] Periyasamy, R., and Nalini, D. (2015). Lloyd Minkowski based K -Means clustering for effective diagnosis of heart disease and stroke. International review on computers and software, 10(6).

[15] Periyasamy, R., and Nalini, D. (2015). Pairwise proximity clustering for medical disease identification. International journal of advanced computer science and technology, 5(1), pp. 1-10.

[16] Periyasamy, R., and Nalini, D. (2015). A clinical heart disease decision supportive optimized mining method for effective disease diagnosis. International journal of computer applications, 126 (9).

[17] Subhashini, M., & Gopinath, R. (2020). Mapreduce Methodology for Elliptical Curve Discrete Logarithmic Problems – Securing Telecom Networks, International Journal of Electrical Engineering and Technology, 11(9), 261-273.

[18] Upendran, V., & Gopinath, R. (2020). Feature Selection Based on Multi criteria Decision Making for Intrusion Detection System. International Journal of Electrical Engineering and Technology, 11(5), 217-226.

[19] Upendran, V., & Gopinath, R. (2020). Optimization Based Classification Technique for Intrusion Detection System. International Journal of Advanced Research in Engineering and Technology, 11(9), 1255-1262.

[20] Jayasimman, L., Geetha Dhanalakshmi, V. (2020). Design of Enhanced Dynamic Resource Allocation Framework for Heterogeneous Cloud Environment, IJITEE, 9(3), 1563-1568.

[21] Jayasimman, L., Geetha Dhanalakshmi, V. (2018). A Study on Spatial-Temporal Load Balancing Approach in Cloud Computing, JCSE, 6(11).

[22] Geetha Dhanalakshmi, V., Jayasimman, L. (2018). A Review on Dynamic Resource Allocation strategies for Cloud Computing, IJSRCSAMS, 7(4).

[23] Geetha Dhanalakshmi, V., Jayasimman, L. (2018). Resource Provisioning for Ensuring QoS in Virtualized Environments, JCSE, 6(4).

[24] James Manoharan, J., Hari Ganesh, S. (2016). A framework for enhancing the efficiency of k-means clustering algorithm to avoid formation of empty clusters. Middle-East J. Sci. Res (MEJSR).

[25] James Manoharan, J., Hari Ganesh, S., Sathiaseelan, JGR. (2016). Outlier detection using enhanced k-means clustering algorithm and weight-based center approach, Int. J. Comput. Sci. Mobile Comput, 5(4), 453-464.

[26] James Manoharan, J., Hari Ganesh, S. (2016). INITIALIZATION OF OPTIMIZED K-MEANS CENTROIDS USING DIVIDE-AND-CONQUER METHOD, ARPN Journal of Engineering and Applied Sciences, 11(2), 1086-1091.

[27] James Manoharan, J., Ganesh, S. H., Felciah, M. L. P., & Shafreenbanu, A. K. (2014, February). Discovering Students' Academic Performance Based on GPA Using K-Means Clustering Algorithm. In 2014 World Congress on Computing and Communication Technologies (pp. 200-202). IEEE.

[28] Vijilesh, VG., Hari Ganesh, S., James Manoharan, J. (2018). An Enhancing the Performance of High Utility Itemset Mining using Utility Information Record, International Journal of Pure and Applied Mathematics, 118(17), 257-272.

[29] Selvaramalakshmi, P., Hari Ganesh, S., James Manoharan, J. (2017). A Novel PSS Stemmer for String Similarity Joins, 2017 World Congress on Computing and Communication Technologies (WCCCT), 147-150, IEEE.