International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST20748

Heart Disease Prediction Based on an Optimal Feature Selection Method using Autoencoder

Azhar M.A., Princy Ann Thomas.

PG Scholar, Department of Computer Science and Engineering, Government Engineering College, Idukki, Kerala, India

Associate Professor, Department of Computer Science and Engineering, Government Engineering College, Idukki, India

Article Info
Volume 7, Issue 4
Page Number: 25-38
Publication Issue :
July-August-2020

Article History
Accepted : 01 July 2020
Published : 11 July 2020

ABSTRACT

Heart Failure is one of the common diseases that can lead to dangerous situations. There are several data available within the healthcare systems. However, there was an absence of successful analysis methods to find connections and patterns in health care data. Some Machine learning methods can help us remedy this circumstance. This helps in getting a better insight into the concept of a classification problem. In many classification problems, it is difficult to learn good classifiers before removing these unwanted features due to the huge size of the data. In my work, we have used an artificial neural network-based autoencoder for effective feature selection The aim of feature selection is improving prediction performance and providing a better understanding of the process data. Hybrid Classification method with a dynamic integration algorithm for classification that aims at finding optimal features by applying machine learning techniques resulting in improving the performance in the prediction of cardiovascular disease.

Keywords : Data Mining, Autoencoder, Hybrid, Classification Model, Dynamic Integration Algorithm

I. INTRODUCTION

Heart disease is one of the most critical human diseases in the world and over the last decade heart disease is the main reason for death in the world. To lower the number of deaths from heart diseases, there has to be a fast and efficient detection technique [1]. The hybrid method is one of the effective data mining methods until this date [2][3]. One of the major challenges in heart disease is correct detection and finding the presence of it inside a human. [4].Machine learning could be a better choice for achieving high accuracy for predicting not only heart disease but also, another disease because the tool utilizes feature vector and it is various data types under various condition for predicting the heart disease, algorithms such as support vector machine and multilayer perceptron, are used to predicate risk

of heart diseases [4][5]. This prediction system for heart disease helps specialists to predict heart disease in the early stage of disease resulting in saving a large number of lives. This paper also does a deep analysis of the utilization of learning in the field of predicting heart disease [6].In the previous study; we included a hybrid feature selection model that produces an enhanced performance level with high accuracy. This model included a Random forest with a Linear model. After the pre-processing Constructed, a Decision Tree on a given dataset and construct, a subset of features. Our result reveals that we can agree with feature selection, but there is no consistent method over the feature selection because we are not in a position to go through all possible subsets. We cannot search for all these subsets to get the optimal one. The other side is supposed we do not search for all possibilities then we will not get the optimal subset. There exist no feature selection algorithm which provides an optimal subset of features for any criterion function without doing exhaustive search where Some criterion function satisfies some properties then we can use those properties to obtain a feature selection algorithm that will give us optimal feature subset without doing an exhaustive search. If an algorithm does not use the properties of criterion function without doing exhaustive search then we cannot guarantee optimality. So we introduce neural network-based autoencoder it solves the problems in a particular extend.

The objectives of the proposed work are.

- To predict the heart disease
- To improve the performance measures of heart disease prediction
- Study the existing risk assessment systems and their limitations and demonstrate the advantage of using nonlinear models for the classification compared to the use of linear models.
- To analyze the best techniques used for heart disease prediction

Our proposed work enhances the performance level by using feature selection using autoencoder and hybrid Classification with a dynamic integration algorithm(DIA).[7].The experiment results show that our proposed hybrid method with DIA has a stronger capability to predict heart disease compared to existing methods. Here this work produces a prediction model using not only distinct techniques but also by relating two or more techniques with a dynamic integration algorithm. These types of combinations of two or more classification algorithms are commonly known as hybrid methods [13]. The dataset UCI is used for classification, where 70% of the data is used for training and the remaining 30% is used for testing [9], [10]. Neural networks and support vector machines are generally regarded as the best tool and classification of diseases. The proposed method which we use has 13 attributes of 76 for heart disease prediction. The results show an enhanced level of performance compared to the previous methods. Their evaluation becomes very important. We generate results using random forest, logistic regression multi-layer perceptron, and support vector machine individually and its hybrid forms, which all produce a good performance in the prediction of heart disease [6], [18]. Neural system strategies are presented, which combine not only posterior probabilities as well as predicted values from multiple predecessor techniques But the hybrid model of multilayer perceptron and support vector machine model achieves an accuracy level of up to 91.17% which is a good accuracy compared to previous works in our experiment environment.

II. RELATED WORKS

In machine learning, different methods used to find risk prediction of heart disease. To Study the

dependencies between the features of the given dataset to understand their influence on prediction accuracy of different machine learning algorithms. Different researches used several techniques to improve the heart disease process such as feature selection and machine learning classifiers [11][12][13]. Therefore, this research attempts to improve the performance of the classifiers by doing experiments using multiple machine-learning models to make better use of the dataset collected from different medical databases. For this, multiple machine learning approaches used to understand the data and predict the heart disease chances in a medical database. Moreover, the results and relative study demonstrated that the flow work improved the past accuracy score in foreseeing heart disease. The integration of the machine learning model presented in this study with medical information systems would be useful to predict heart disease or any other disease using the data collected from patients. The number of features is reduced from 13 to 7 and the accuracy of the proposed method is 86%. Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm is presented in [14] [15]. Another model combines a neural network, fuzzy logic, and genetic algorithm. Implemented a system for predicting heart disease using Data mining techniques: K Means and Weighted Association rule. Results demonstrate that K-means with decision tree technique makes the system more accurate and efficient compared to the weighted association rule with the Apriori algorithm. Heart disease prediction system using three data mining classification techniques (Decision Trees, Naive Bayes, Neural Networks), two more attributes: Obesity and smoking along with consistent 13 attributes are used in [16]. J48 decision tree algorithm which uses a pruning method for building a tree and data mining tool Weka 3.6.6 is used. Results show that neural systems produce accurate results by comparing the accuracy of classification strategies. Multi-layer Perceptron Neural Networks (MLPNN) is used for improving accuracy. Artificial neural networks with backpropagation error method are used to classify the cerebrovascular disease [17]. The neural network was trained with 13 input attributes using a backpropagation algorithm with sigmoid function on one hidden layer which improves the accuracy. Many studies exist to predict heart disease with various techniques. Different researches used several techniques to improve the heart disease process such as feature selection and machine learning classifiers So it is necessary to build an efficient intelligent trusted automated system that predicts the heart disease accurately based on the symptoms according to gender/age and domain knowledge of experts in the field at the lowest cost[18]. The previous hybrid approach is used combining the HRFLM characteristics of Random Forest (RF) and Linear Method (LM) proved to be quite accurate in the prediction of HRFLM heart disease. The model which is proposed for Heart attack Prediction System is invented for using traditional algorithms and approaches. But by using all the existing systems the accuracy is very less.

III. PROPOSED METHOD

In this study (see Fig. 1) network-based Autoencoder was used to select important features, and on these selected features, the performance of the classifiers (hybrid SVM and MLP) was tested and compares the performance measures. After the Comparision of performance, we can understand that the Autoencoder gets the optimum features, and it shows the best classification result. Classifiers logistic regression, MLP, SVM, RF, and its hybrid combination with DIA were used in the system.. The methodology of the proposed system structured into five stages including(fig 1) (A) pre-processing of the dataset, (B) feature selection, (C) classification,(D)hybrid with DIA approaches (E) Dynamic integration algorithm, and (E)classifier's

performance evaluation F)Experiments and Results. In this work, the medical data related to Heart diseases are considered. This benchmark dataset was obtained from the Cleveland UCI repository [19]. This is a publicly available dataset. Cleveland dataset concerns the classification of a person into a no cardiac and cardiac person regarding heart diseases. The dataset is divided into 2 sets of training (70%) and testing (30%). Python is used for the experiment. The data consists of 13 attributes (inputs) and 2 classes (outputs). Tables 1 illustrate the representation of the attributes.

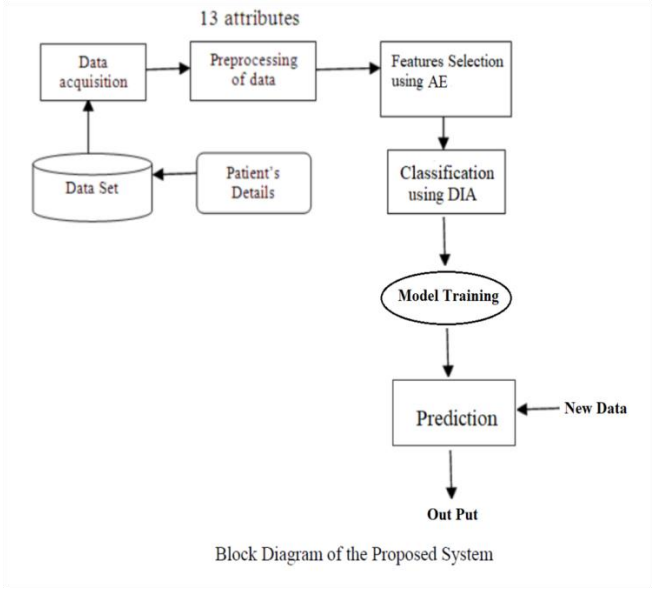


Fig 1: Block diagram of the proposed model

A. *Data set pre-processing*

Pre-processing is a strategy that is utilized to convert the raw data into a perfect informational collection. After the data collection ML process starts from a pre-processing data phase followed by feature selection based on autoencoder. Table.1 shows the UCI dataset detailed information with attributes used. For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some predetermined Machine Learning model needs data in a specified format. Data Preprocessing is necessary for efficient representation of data and machine learning classifiers which should be trained and tested effectively. In particular, the Cleveland and Hungarian databases have been used by many researchers and found to be suitable for developing a mining model, because of lesser missing values and outliers.

B. *Feature selection and Reduction*

Feature selection is a procedure in machine learning to find a subset of features, it removes irrelevant and redundant information as much as possible. The objective of feature selection is to improve the prediction performance and providing a better understanding of the process data. We also review some of the feature selection techniques on standard datasets to illustrate the applicability of feature selection techniques. we used the artificial neural network-based autoencoder, which aims to the feature selection problem more than all other conventional feature selection methods. Here we trained autoencoder network with relu and sigmoid being the activation functions for screening out the redundancy. We can observe that the autoencoder methods were more stable than the other individual technique.

TABLE 1. UCI dataset attributes detailed information.

Si.No	Description
1	**Age** - defining the age of the person. [Minimum Age: 29, Maximum Age: 77]
2	**Sex** - [—0‖ means Female and —1‖ means Male]
3	**CP** - is defining the level of chest pain (CP) a patient suffering from, when reached the hospital. There are four kinds of distinct values defined for this attribute, where each value is describing a level of chest pain.
4	**trestBP** - describes the blood pressure (BP) figure for the patient while admitted to the hospital. [Minimum BP: 94, Maximum BP: 200]

5	**Chol** - showing the cholesterol level recorded while admitting the patient in the hospital. [Minimum Chol: 126, Maximum Chol: 564]
6	**FBS** - is describing the fasting blood sugar level in the patient. It has binary classified values. The values are depending on, if the patient has more than 120mg/dl sugar = 1, if not = 0.
7	**restECG** - is showing the result of ECG from 0 to 2. Where each value is showing the severity of the pain.
8	**Thalach**(HeartBeat) - The maximum value of heartbeat counted at the time of admission [Minimum: 71, Maximum: 202]
9	**Exang** - understand about, does exercise induce angina or not. If yes, the value will be —1‖, and —0‖ for not.
10	**OldPeak** - is defining the patient's depression status. It is assigned as different real number values fall between 0 and 6.2.
11	**Slope** - The condition of the patient during peak exercise. This value defined into three segments [Upsloping, Flat, Down sloping]
12	**CA**: is showing the status of fluoroscopy. It is showing that how many vessels are colored.
13	**Thal** - is another kind of test required for the patient having chest pain or breathing difficulty. Four kinds of values showing the result of Thallium test.

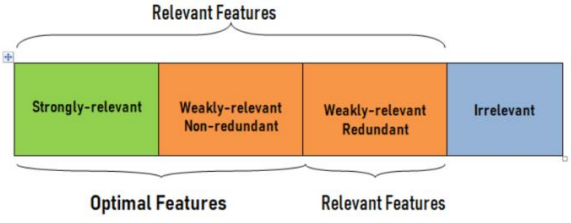


Fig 2: features from a dataset

Autoencoder: A feature selector: a neural network that the output is the input itself. The traditional autoencoder is an artificial neural network that attempts to reproduce its input, i.e., the target output is the input. Autoencoders are composed of an input, a hidden, and an output layer (see fig 3). Autoencoders get familiar with a "compressed representation" of input automatically by first compressing the input (encoder) and decompressing it back (decoder) to match the original input.". The learning is aided by using distance function that quantifies the information loss that occurs from the lossy compression. Inside, it has a hidden layer 'h' that portrays a code used to represent the input.

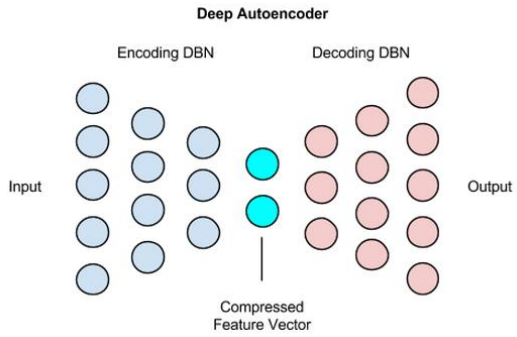


Fig 3: Autoencoder

C. *Classification Modelling*

After analysis, we had concluded that SVM and MLP are better than other methods for the prediction of heart disease. these methods are discussed bellow

1)Support Vector Machine (SVM)

From the below analysis, Sequential minimal optimization in the Support vector machine is more effective. The support vectors are the data coordinates that are on the boundary of the margin. Mathematical functions are involved in SVM design which is frequently used to model real-world problems when we have large data set of entries The system with SVM will help for early diagnosis of heart disease for any given patient. This system when deployed can

complement traditional heart disease detection systems and can help not only the doctors but also the patients. The dataset with 13 attributes is fed to the SVM classifier and the output is mapped into two classes cardiac and noncardiac. SVM aims to identify the best classification function to distinguish between members of the two classes from the training data. The classification performance of the heart disease dataset is discussed in the following section.

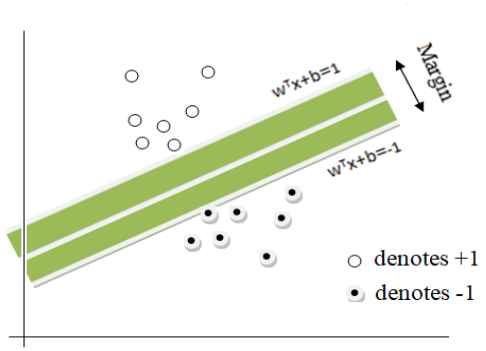


Fig 4: The SVM classification of points

SVM is one of the most effective classifiers among those which are sought of linear. It has a very good mathematical intuition behind the SVM and we can handle certain cases where there is on linearity by using nonlinear basis function is called kernel function.SVM has a clever way to prevent overfitting and we can work with a relatively larger number of features without requiring computation.

2)Artificial neural network(ANN)

Artificial neural network models are a first-order mathematical approximation to the human nervous system that has been widely used to solve various nonlinear problems. It is a "connectionist" computational system. A true neural network does not follow a linear path or it exploits the nonlinearity. Every layer of the NN computes a nonlinear function of the input feature vector. The input to the particular layer of neurons and the output of that layer of the neuron is an intermediate feature vector of the same or different dimension depending upon how many neurons have in that particular layer. One of the key elements of a neural network is its ability to learn and it can change its internal structure based on the information flowing through it.we have a Perceptron convergence algorithm.

1. If the output unit is correct to leave its weights alone.
2. If the output unit is incorrect outputs a zero, add the input vector to the weight vector
3. If the output unit is incorrect outputs a one, subtract the input vector to the weight

Vectorization of perceptron model:

$$\begin{bmatrix} x_0 \\ x_1 \\ . \\ . \\ x_n \end{bmatrix} \text{X } [w_0 \quad w_1 \,..\, w_n] = x_0 w_0 + x_1 w_1 + .., + x_n w_n$$

$$y = \sum_{i=0}^{n} x_i w_i + b$$

$$= w^T.x$$

b=w0x0,where x0=1

Multi-Layer Perceptron: Every layer of neuron computes a nonlinear function and we have a cascade of the nonlinear function that is in f (i+1) th layer it computes the nonlinear function of $f^{(i)}$ on the feature vector which are computed by previous layer $f^{(i-1)th}$ similarly we can cascade all of them together. In the system for heart disease prediction, the multilayer perceptron architecture of the neural network is used. The system consists of two steps, in the first step 13 clinical attributes are accepted as input and then the training of the network is done with training data by the back-propagation learning algorithm. Training these networks is much more complicated. Multilayer is feed-forward neural networks trained with the standard back-propagation algorithm.

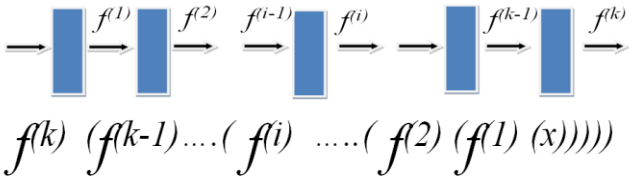


Fig 5: Mathematical illustration of MLP

Training the network (i.e. adjusting the weights) also involves taking the error (desired result - guess). The error, however, must be fed back through the network. The final error ultimately adjusts the weights of all the connections The multilayer is trained with error-correction learning, which is appropriate here because the desired multilayer response is the arteriographic result and as such known. Error correction learning works in the following way from the system response at neuron j at iteration t, $yj(t)$, and the desired response $dj(t)$for given input pattern an instantaneous error $ej(t)$is defined by $ej(t) = dj(t) - yj(t)$ (1) Using the theory of gradient descent learning, each weight in the network can be adapted by correcting the present value of the weight with a term that is proportional to the present input and error at the weight, i.e. $wjk(t + 1) = wjk(t) + \eta\delta j(t)xk(t)$ (2) The $\eta(t)$ is the learning rate parameter. The $wjk(t)$ is the weight connecting the output of neuron k to the input neuron j at iteration t. The local error$\delta j(t)$ can be computed as a weighted sum of errors at the internal neurons. The BP algorithm has served as a useful methodology to train a multilayer perceptron for a wide range of applications. The BP network calculates the difference between real and predicted values, which is circulated from output nodes backward to nodes in the previous layer.

D. *Hybrid with DIA Approaches*

With hybrid architecture, at least two models are coordinated to get a cooperative effect where the strengths of one system can compensate for the shortcoming of another. For the perfect analysis of heart disease, the outputs of each algorithm are combined and compared. Applying hybrid data mining strategies can show promising results in the finding of heart disease. The Algorithm1 shows the Proposed Approach of the Hybrid Data Mining Technique. After the pre-processing Construct subset of features using autoencoder. Then, the classifiers SVM and MLP are applied to the selected dataset to estimate its performance. Table 4-shows a Comparison of various models with the hybrid model.

Hybrid Algorithm 1

Step-1: Pre-process the data set to remove the duplicate,
missing and unknown data.
Step-2: Construct an Autoencoder neural network on given
Dataset.
Step-3: Input all features to the Autoencoder network and
create optimum features.
Step-4: Learn and Classify the Hybrid Model (SVM + MLP)
with selected Features
Step-5: input step 4 to the DIA algorithm (Call Algorithm 2).

E. *Dynamic Integration Algorithms*

In these techniques, we performed experiments on the training datasets and the results are collected. For approximately 97% of the test samples, the prediction score above 0.6 and below 0.4 is the correct score for training datasets. Also, the majority of the misclassification lies in the range of 0.4 and 0.6. For training data, many misclassified samples lie in between the range of 0.3- 0.6. However, we choose not to use 0.3 as the lower threshold because the range between 0.3 to 0.6 contains much more correct predictions than wrong predictions. Hence, including them in the decision-making process will reduce the overall accuracy. In this way, only scores that lie in

the middle of 0.4 - 0.6 would be considered for dynamic decision making. We then identified the higher scoring classifiers based on the features. The classifier is noted for that particular dataset and it is used as a starting selection point in our decision-making module as illustrated in DI Algorithm 2.We already know the high scoring classifier, we pick the initial prediction answer from the high scoring classifier saw during the tests. At this stage, we introduce a new parameter, δ. The purpose of δ is to decide if we should discard the prediction score of the previously selected high scoring classifier. The selection of δ was done after performing experiments with multiple values of δ. The best score over all the sensors in the database of UCI was produced with δ as 0.2. After getting the absolute difference between the two scores, we compare it with the δ. If the difference in the score is less than the value of δ, we select the final result from the earlier selected high scoring classifier. If it is more than delta, we select the result from the lower-scoring classifier.

Dynamic Integration Algorithm 2
Input: scoreLow (i), scoreHigh (i), testLabel (i), N;
Ouput: finalScore (i);

```
Initialization,finalScore,diffInScore;
For i =0 to N do
    finalScore= scoreHigh(i);
    diffInScore=abs(scoreHihg(i)-ScoreLow(i));
    if scoreHigh(i) ≥ 0.4 and scoreHigh(i) ≤ 0.6 then
        if diffInScore ≤ 0.2 then
            finalScore= scoreHigh(i);
        else
            finalScore= scoreLow(i);
        end
    If finalScore ≤ 0.5 then
        result=0
    else
        result=1
    end
end
```

IV. PERFORMANCE MEASURES

Several standard performance measures such as accuracy, precision, and error in classification have been considered for the computation of the performance efficiency of this model. To check the performance of the classifiers, various performance evaluation metrics were used in this project work. We used the confusion matrix and ROC for evaluation.

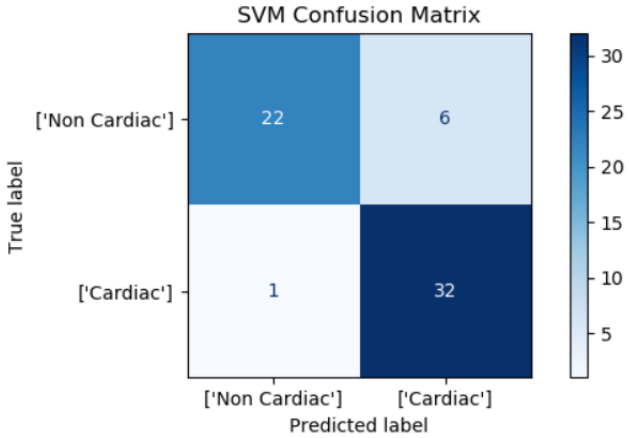


Fig 6: confusion matrix of SVM

Confusion Matrix: Here, every observation in the testing set is predicted in exactly one box. With just two classes, there are four possible results with the classification. The upper left and lower right quadrants represent the correct actions. The remaining two quadrants are incorrect actions. The performance of classification could be determined by associating costs with each of the quadrants. Classification accuracy is normally determined by deciding the percentage of tuples placed in the correct class. Figure 14. Shows the confusion matrix off SVM and MLP.

Receiver operating characteristic curve (ROC): ROC curve shows the relationship between false positives and true positives. A ROC curve was originally used in the communications area to examine false alarm rates. It has also been used in information retrieval to examine fallout (percentage of retrieved that are not relevant) versus recall (percentage of retrieved that

are relevant). At the beginning of evaluating a sample, there is none of either category, while at the end there is 100 percent of each. Figure 15. shows each new tuple is either a false positive or true positive.

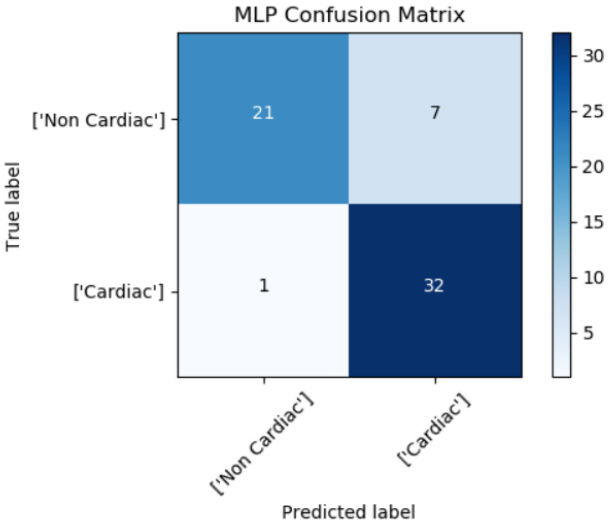


Fig 7: confusion matrix of MLP

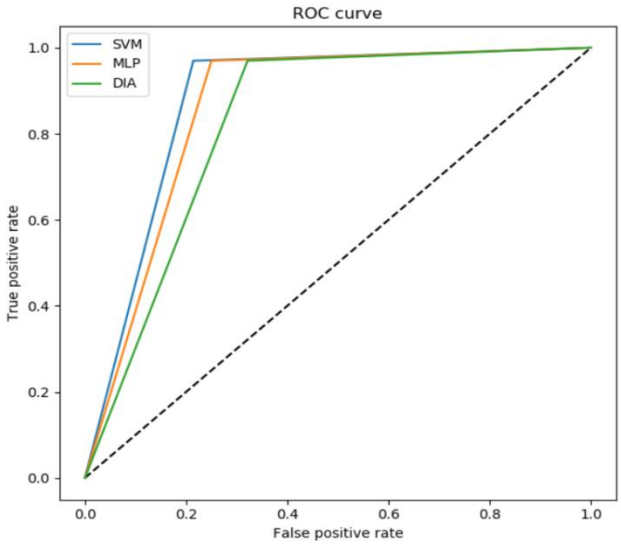
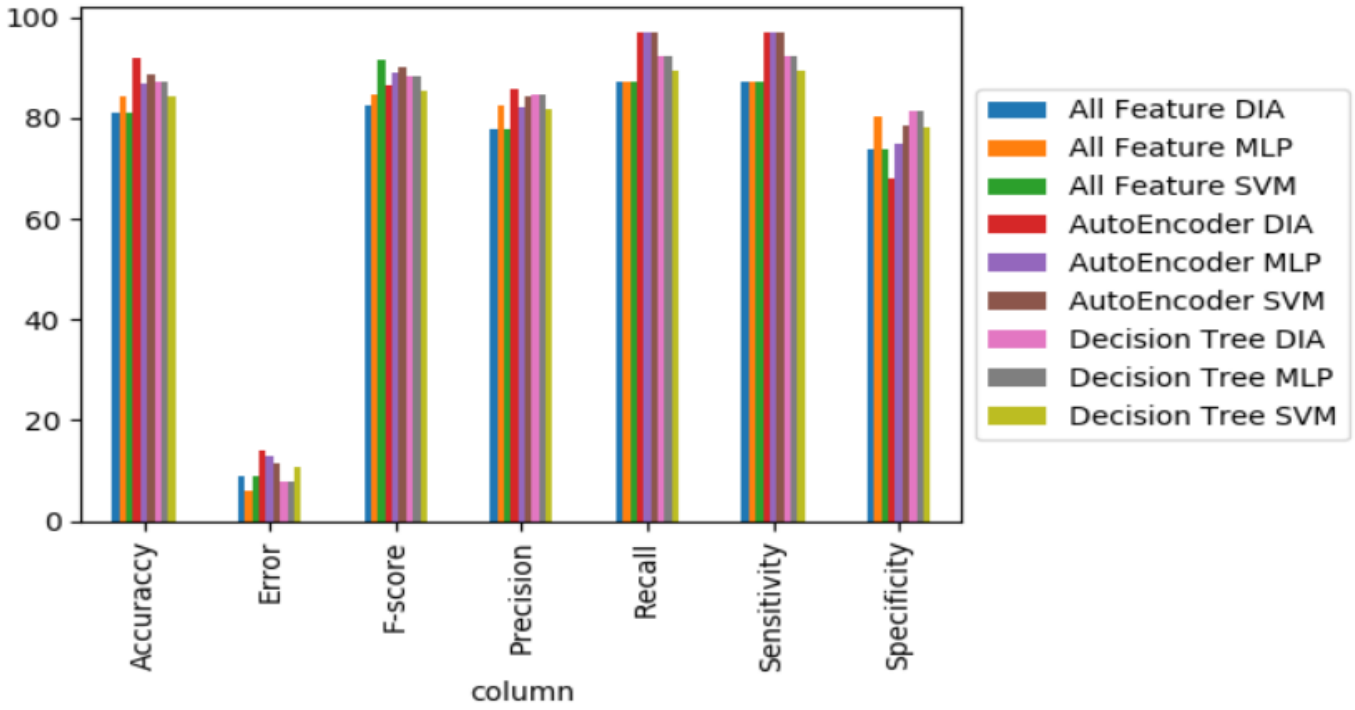
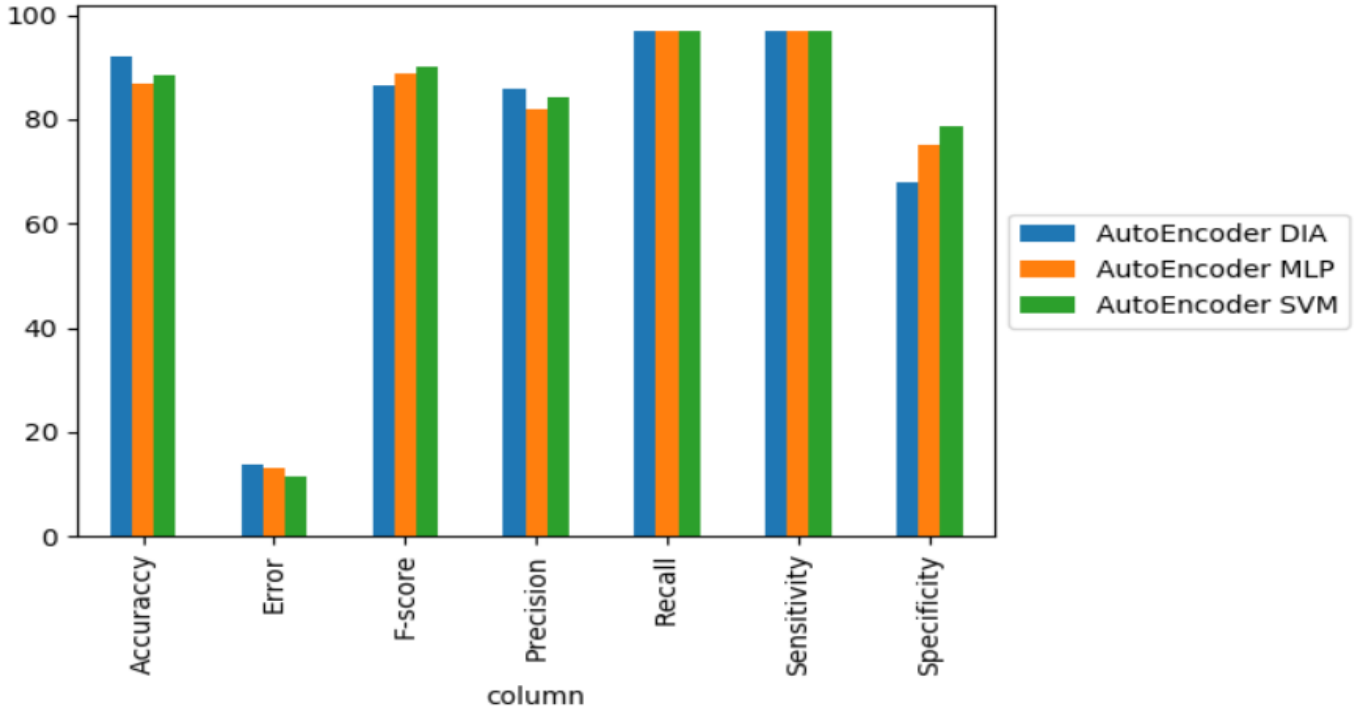


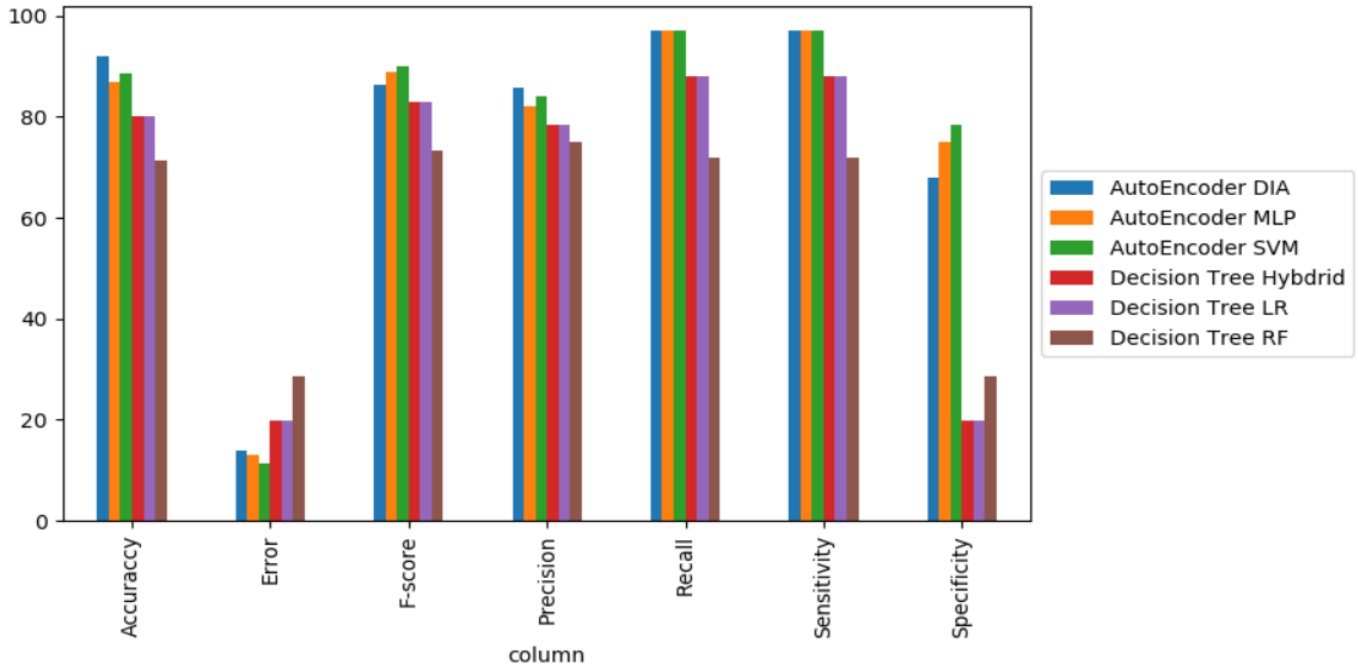
Fig 8: ROC curve

TABLE 2. Result of various models with the proposed model.

Classification Methods	Feature selection Methods	Accuracy	Precision	Sensitivity	Specificity	F-Score	Error
SVM	All features	81.5	77.84	87.27	73.93	**91.43**	8.85
MLP		84.1	82.29	87.27	80.36	84.71	**5.9**
(SVM +MLP) with DIA		81.15	77.84	87.27	73.93	82.29	8.85
SVM	Decision tree	84.1	81.81	89.24	78.04	85.36	10.9
MLP		87.21	84.44	92.12	81.43	88.12	7.79
(SVM +MLP)with DIA		87.21	84.44	92.12	**81.43**	88.12	7.79
SVM	Proposed AE method	88.52	84.21	96.97	78.57	90.14	11.48
MLP		86.89	82.05	96.97	75.0	88.89	13.11
(SVM +MLP)with DIA		**91.97**	**85.85**	**96.97**	67.86	86.49	13.93


100
80
60
40
20
0
Accuraccy
Error
F-score
Precision
Recall
Sensitivity
Specificity
column
All Feature DIA
All Feature MLP
All Feature SVM
AutoEncoder DIA
AutoEncoder MLP
AutoEncoder SVM
Decision Tree DIA
Decision Tree MLP
Decision Tree SVM


100
80
60
40
20
0
Accuraccy
Error
F-score
Precision
Recall
Sensitivity
Specificity
column
AutoEncoder DIA
AutoEncoder MLP
AutoEncoder SVM



V. EXPERIMENT AND RESULTS

This section of the paper summarises the discussion on the classification models and their outcomes from different perspectives. First, we checked the performance of the following machine learning algorithms SVM, MLP, and hybrid model contains SVM and MLP by dynamic integration on full features. In the second, we used a decision tree feature selection algorithm for important feature subset selection and apply these subsets of features to the above-mentioned machine learning algorithms. In the third, performances were checked on selected features using artificial neural network-based autoencoder. To check the performance of classifiers performance evaluation metrics were applied. The experimental results in Table 7 show that the selected set of attributes using an autoencoder, hybrid with a dynamic integration algorithm, and the appropriate training set gives a better performance when compared to the existing method. Next, we get the performance of the machine learning algorithm LR, RF, and hybrid(LR+RF) with include feature subset selection using decision tree and get the presentation of the machine learning algorithm SVM, MLP, and hybrid model contains SVM and MLP by dynamic integration with highlight features from autoencoder. If interrelationship between two features cannot be mined effectively then a single method is not efficient. Autoencoder based feature selection tries to extract this interrelationship between features and then thereby improves the performance of the Machine Learning algorithm. After the examination, we can say that hybrid classifiers with DIA show more Accuracy than normal hybrid and all other Individual methods.

VI. CONCLUSION AND FUTURE WORK

This work will be useful in identifying possible patients who may suffer from heart disease in the coming years. The comparison of the accuracy rates of all classifiers can be seen in Table 7. Overall, there is little fluctuation in the accuracy rates of all classifiers. Inter-Related feature selection can outperform traditional feature selection methods. Hybrid classifiers with DIA shows the best performance measures, their accuracy rates are the highest compared to other techniques. Here we use a network-based Autoencoder feature selection technique. The results also demonstrate that the

reduced feature subset can have better prediction performance compared to the original set of attributes. Our proposed system tries to extract these interrelationships between features and select the best feature subset and thereby improve the performance of the Machine Learning algorithm. The performance of feature selection techniques depends on the type of dataset that we have taken for experimenting. Also, we can observe that the hybrid methods with DIA were more stable than the other individual techniques. The results indicated that the system can be useful and helpful for the doctors for timely diagnoses the chances of a heart attack in a patient.

There are some shortcomings to be a further study in the future. The focus of the latter research will be on the other data set to verify its classification performance. At the same time, our training sample data is too small, it is necessary to increase more data sets to test better. As far as UCI dataset concerns, the dataset needs to be amplified.

VII. REFERENCES

[1]. Jabbar MA, Chandra P, Deekshatulu BL. Cluster-based association rule mining for heart attack prediction. Journal of Theoretical and Applied Information Technology. 2011; 32(2):197–201.

[2]. Sudha A, Gayathiri P, Jaisankar N. Effective analysis and predictive model of stroke disease using classification methods. International Journal of Computer Applications. 2012; 43(14):26–31.

[3]. Amin SU, Agarwal K, Beg R. Genetic neural network-based data mining in the prediction of heart disease using risk factor. Proceeding of IEEE Conference on Information and Communication Technologies (ICT); 2013 Apr. p. 1227– 31.

[4]. Deepika N, Chandrashekar K. Association rule for classification of Heart Attack Patients. International Journal of Advanced Engineering Science and Technologies. 2011; 11(2):253–57.

[5]. Sellappan Palaniappan and Rafiah Awang (2008): Intelligent Heart Disease Prediction System Using Data Mining Techniques, 978-1-4244-1968- 5/08/ IEEE.

[6]. Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm M. ANBARASI, E. ANUPRIYA,N.CH.S.N.IYENGAR.-2012

[7]. Feature Selection using Artificial Bee Colony for Cardiovascular Disease Classification B.Subanya, Dr.R.R.Rajalaxmi-2014

[8]. "Analysis of data mining techniques for heart disease prediction," 2016-2017, M. Sultana, A. Haider, and M. S. Uddin

[9]. D. K. Srivastava and L. Bhambhu, "Data classification using support vector machine," J.Theor. Appl. Inf. Technol., 2009

[10]. N. Bhatia and C. Author, "Survey of Nearest Neighbor Techniques," IJCSIS) Int. J. Comput. Sci. Inf. Secur., vol. 8, no. 2, pp. 302–305, 2010.

[11]. T. M. Lakshmi, A. Martin, R. M. Begum, and V. P.Venkatesan, "An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data," Int. J. Mod. Educ. Comput. Sci., vol. 5, no. 5, pp. 18–27, 2013.

[12]. Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data Simon Fong, Raymond Wong, and Athanasios V. Vasilakos, Senior Member, IEEE

[13]. T.Peter and K.Sonausundaram "An empirical study on prediction of heart disease using classification data minings techniques" in IEEE International Conference and Management-2012

[14]. A. Khemphila and V.Boonjing "Comparing Performance of logistic regression decision trees and neural networks for classifying heart disease patients "in International Conference on Computer Information systems and Industrial Management Applications"

[15].Data Mining Techniques on Risk Prediction: Heart Disease, G. Purusothaman* and P. Krishnakumari, Indian Journal of Science and Technology-2015

[16].A Fuzzy Rule-based Approach to Predict Risk Level of Heart Disease, By Kantesh Kumar Oad & Xu Delhi-2014

[17].The Best Two Independent Measurements Are Not the Two Best"-1974, THOMAS. M. COVER

[18].Enhanced Prediction of Heart Disease by Genetic Algorithm and RBF Network-2015, A. Durga Devi

[19].E. J. Benjamin, P. Muntner, and et al. Alonso, Alvaro, ―Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association,‖ Circulation, vol. 139, no. 10, 2019.

[20].M. Ramaraj and T. A. Selvadoss, ―A Comparative Study of CN2 Rule and SVM Algorithm and Prediction of Heart Disease Datasets Using Clustering Algorithms,‖ Netw. Complex Syst., vol. 3, no. 10, pp. 1–6, 2013.

[21].A. Gavhane, G. Kokkula, I. Pandya, and P. K. Devadkar, ―Prediction of Heart Disease Using Machine Learning,‖ in Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology, ICECA 2018, 2018, pp. 1275–1278.

[22].C.-S. Lee and M.-H. Wang, ―A fuzzy expert system for diabetes decision support application.,‖ IEEE Trans. Syst. MAN, Cybern. B Cybern., vol. 41, no. 1, pp. 139–153, 2011.

[23].C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, ―Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field,‖ in Machine Learning Paradigms, 2019, pp. 71–99.

[24].K. Shameer, K. W. Johnson, B. S. Glicksberg, J. T. Dudley, and P. P. Sengupta, ―Machine learning in cardiovascular medicine: are we there yet? ‖ Heart, vol. 104, no. 14, pp. 1156–1164, 2018.

[25].A. M. Shahiri and W. Husain, "A review on predicting student's performance using data mining techniques," Procedia Computer Science, vol. 72, pp. 414-422, 2015.

[26].C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, pp. 601-618, 2010.

[27].M. Ramaswami and R. Bhaskaran, "A study on feature selection techniques in educational data mining," arXiv preprint ar X iv:0912.3924, 2009.

[28].A. Mueen, B. Zafar, and U. Manzoor, "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques," International Journal of Modern Education and Computer Science, vol. 8, p. 36, 2016.

[29].W. Punlumjeak and N. Rachburee, "A comparative study of feature selection techniques for classify student performance," in Information Technology and Electrical Engineering (ICITEE), 2015 7th International Conference on, 2015, pp. 425-429.

[30].D. Koller and M. Sahami, "Toward optimal feature selection," Stanford InfoLab1996.

[31].P. Mitra, C. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," IEEE transactions on pattern analysis and machine intelligence, vol. 24, pp. 301-312, 2002.

[32].H. M. Harb and M. A. Moustafa, "Selecting an optimal subset of features for student performance model," Int J Comput Sci, p. 5,2012.

[33].A. Figueira, "Predicting Grades by Principal Component Analysis: A Data Mining Approach to Learning Analytics," in Advanced Learning Technologies (ICALT), 2016 IEEE 16th International Conference on, 2016, pp. 465-467.

[34].E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining Educational Data to Predict Student's academic Performance using Ensemble Methods," International Journal of Database

Theory and Application, vol. 9, pp. 119-136, 2016.

[35].K. Patel, J. Vala, and J. Pandya, "Comparison of various classification algorithms on iris datasets using WEKA," Int. J.Adv. Eng. Res. Dev. (IJAERD), vol. 1, 2014.

[36].M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update,"ACM SIGKDD explorations newsletter, vol. 11, pp. 10-18,2009.

Webliography

[37].https://en.wikipedia.org/wiki/Feature_selection#/media/File: created on 7 march 2019 and accessed on 13 September 2019

[38].https://web.stanford.edu › class › handouts › CS276: Information Retrieval and Web Search Christopher Manning and Pandu Nayak accessed on 26 September 2019

[39].https://images.app.goo.gl/7SGi2Y6sHwwiFGNa9 created on 13 march 2029 and accessed on 10 April 2020

Cite this article as :

Azhar M. A., Princy Ann Thomas, "Heart Disease Prediction Based on an Optimal Feature Selection Method using Autoencoder", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 7 Issue 4, pp. 25-38, July-August 2020. Available at doi :
https://doi.org/10.32628/IJSRST20748
Journal URL : http://ijsrst.com/IJSRST20748