EFFECTIVE HEART DISEASE PREDICTION USING HYBRID MACHINE LEARNING TECHNIQUES

Gundam Poojitha(21645A6708), B.tech Student, CSE(Data Science), Vaagdevi College of Engineering

Ponnala Bharath(20641A6753), B.tech Student, CSE(Data Science), Vaagdevi College of Engineering

Chiguru Keerthi Dharani(20641A6718), B.tech Student, CSE(Data Science), Vaagdevi College of Engineering

Mohammed Inkeshaf Aliuddin(21645A6711), B.tech Student, CSE(Data Science), Vaagdevi College of Engineering

Mrs.R.Divija, Assistant professor, Department of CSD, Vaagdevi College of Engineering

ABSTRACT

Heart disease is one of the most significant causes of mortality in the world today. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis. Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. We have also seen ML techniques being used in recent developments in different areas of the Internet of Things (IoT). Various studies give only a glimpse into predicting heart disease with ML techniques. In this paper, we propose a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques. We produce an enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM)

1.INTRODUCTION

It is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Various techniques in data mining and neural networks have been employed to find out the severity of heart disease among humans. The severity of the disease is classified based on various methods like K-Nearest Neighbour Algorithm (KNN), Decision Trees (DT), Genetic algorithm (GA), and Naive Bayes (NB) [1]. The nature of heart disease is complex and hence, the disease must be handled carefully. Not doing so may affect the heart or cause premature death. The perspective of medical science and data mining are used for discovering various sorts of metabolic syndromes. Data mining with classification plays a significant role in the prediction of heart disease and data investigation. We have also seen decision trees be used in predicting the accuracy of events related to heart disease. Various methods have been used for knowledge abstraction by using known methods of data mining for prediction of heart disease. In this work, numerous readings have been carried out to produce a prediction model using not only distinct techniques but also by relating two or more techniques. These amalgamated new techniques are commonly known as hybrid methods. We introduce neural networks using heart rate time series. This method uses various clinical records for prediction such as Left bundle branch block (LBBB), Right bundle branch block (RBBB) [2] , Atrial fibrillation (AFIB), Normal Sinus Rhythm (NSR), Sinus bradycardia (SBR), Atrial flutter (AFL), Premature Ventricular Contraction (PVC)), and Second degree block (BII) to find out the exact condition of the patient in relation to heart disease. The dataset with a radial basis function network (RBFN) is used for classification, where 70% of the data is used for training and the remaining 30% is used for classification. We also introduce Computer Aided Decision Support System (CADSS) [3] in the field of medicine and research. In previous work, the usage of data mining techniques in the healthcare industry has been shown to take less time for the prediction of disease with more accurate results. We propose the diagnosis of heart disease using the GA. This method uses effective association rules inferred with the GA for tournament selection, crossover and the mutation which results in the new proposed fitness function. For experimental validation, we use the well-known Cleveland dataset which is collected from a UCI machine learning repository. We will see later on how our results prove to be prominent when compared to some of the known supervised learning techniques. The most powerful evolutionary algorithm Particle Swarm Optimization (PSO) [4] is introduced and some rules

are generated for heart disease. The rules have been applied randomly with encoding techniques which result in improvement of the accuracy overall. Heart disease is predicted based on symptoms namely, pulse rate, sex, age, and many others. The ML algorithm with Neural Networks is introduced, whose results are more accurate and reliable as we have seen. Neural networks are generally regarded as the best tool for prediction of diseases like heart disease and brain disease. The proposed method which we use has 13 attributes for heart disease prediction. The results show an enhanced level of performance compared to the existing methods in works. The Carotid Artery Stenting (CAS) has also become a prevalent treatment mode in the medical field during these recent years. The CAS prompts the occurrence of major adverse cardiovascular events (MACE) [5],[6],[7] of heart disease patients that are elderly. Their evaluation becomes very important. We generate results using a Artificial Neural Network ANN, which produces good performance in the prediction of heart disease [6], [15]. Neural network methods are introduced, which combine not only posterior probabilities but also predicted values from multiple predecessor techniques. This model achieves an accuracy level of up to 89.01% which is a strong results compared to previous works. For all experiments, the Cleveland heart dataset is used with a Neural Network NN [8] to improve the performance of heart disease as we have seen previously. We have also seen recent developments in machine learning ML techniques used for Internet of Things (IoT) as well. ML algorithms on network traffic data has been shown to provide accurate identification of IoT devices connected to a network. Meidan et al. collected and labeled network traffic data from nine distinct IoT devices, PCs and smartphones. Using supervised learning, they trained a multi-stage meta classifier[9]. In the first stage, the classifier can distinguish between traffic generated by IoT and non-IoT devices. In the second stage, each IoT device is associated with a specific IoT device class. Deep learning is a promising approach for extracting accurate information from raw sensor data from IoT devices deployed in complex environments. Because of its multilayer structure, deep learning is also appropriate for the edge computing environment. In this work, we introduce a technique we call the Hybrid Random Forest with Linear Model (HRFLM) [10]. The main objective of this research is to improve the performance accuracy of heart disease prediction. Many studies have been conducted that results in restrictions of feature selection for algorithmic use. In contrast, the HRFLM method uses all features without any restrictions of feature selection. Here we conduct experiments used to identify the features of a machine learning algorithm with a hybrid method. The experiment results show that our proposed hybrid method has stronger capability to predict heart disease compared to existing methods.

2.LITERATURE SURVEY

Cardiovascular system diseases are an important health problem. These diseases are very common also responsible for many deaths. With this study, it is aimed to analyze factors that cause Coronary Artery Disease using Random Forests Classifier. According to the analysis, we observed correct classification ratio and performance measure that creates susceptibility to Coronary Artery Disease for each factor. The performance measure results clearly show the impact of demographic characteristics on CAD[11]. Additionally, this study shows that random forests algorithm can be used to the processing and classification of medical data such as CAD.

Heart disease is still a growing global health issue. In the health care system, limiting human experience and expertise in manual diagnosis leads to inaccurate diagnosis, and the information about various illnesses is either inadequate or lacking in accuracy as they are collected from various types of medical equipment. Since the correct prediction of a person's condition is of great importance, equipping medical science with intelligent tools for diagnosing and treating illness can reduce doctors' mistakes and financial losses. In this paper, the Particle Swarm Optimization (PSO)[12] algorithm, which is one of the most powerful evolutionary algorithms, is used to generate rules for heart disease. First the random rules are encoded and then they are optimized based on their accuracy using PSO algorithm. Finally we compare our results with the C4.5 algorithm.

Recently, several software's, tools and various algorithms have been proposed by the researchers for developing effective medical decision support systems. Moreover, new algorithms and new tools are continued to develop and represent day by day. Diagnosing of heart disease is one of the important issue and many researchers investigated to develop intelligent medical decision support systems to improve the ability of the physicians. Neural network is widely used tool for predicting heart disease diagnosis. In this research paper, a heart disease [13] prediction system is developed using neural network. The proposed system used 13 medical attributes for heart disease predictions. The experiments conducted in this work have shown the good performance of the proposed algorithm compared to similar approaches of the state of the art.

The development of medical domain applications has been one of the most active research areas recently. One example of a medical domain application is a detection system for heart

disease based on computer-aided diagnosis methods, where the data is obtained from some other sources and is evaluated by computer based applications. Up to now, computers have usually been used to build knowledge based clinical decision support systems which used the knowledge from medical experts, and transferring this knowledge into computer algorithms was done manually. This process is time consuming and really depends on the medical expert's opinion, which may be subjective. To handle this problem, machine learning techniques have been developed to gain knowledge automatically from examples or raw data. Here, a weighted fuzzy rule-based clinical decision support system (CDSS) [14] is presented for the diagnosis of heart disease, automatically obtaining the knowledge from the patient's clinical data. The proposed clinical decision support system for risk prediction of heart patients consists of two phases, (1) automated approach for generation of weighted fuzzy rules and decision tree rules, and, (2) developing a fuzzy rule-based decision support system. In the first phase, we have used the mining technique, attribute selection and attribute weightage method to obtain the weighted fuzzy rules. Then, the fuzzy system is constructed in accordance with the weighted fuzzy rules and chosen attributes. Finally, the experimentation is carried out on the proposed system using the datasets obtained from the UCI repository and the performance of the system is compared with the neural network-based system utilizing accuracy, sensitivity and specificity.

3. EXISTING SYSTEM

ANN has been introduced to produce the highest accuracy prediction in the medical field. The back propagation multilayer perception (MLP) of ANN is used to predict heart disease. The obtained results are compared with the results of existing models within the same domain and found to be improved . The data of heart disease patients collected from the UCI laboratory is used to discover patterns with NN, DT, Support Vector machines SVM, and Naive Bayes [15]. The results are compared for performance and accuracy with these algorithms. The proposed hybrid method returns results of 86.8% for F-measure, competing with the other existing methods. The classification without segmentation of Convolutional Neural Networks (CNN) is introduced. This method considers the heart cycles with various start positions from the Electrocardiogram (ECG) signals in the training phase.

Disadvantages:

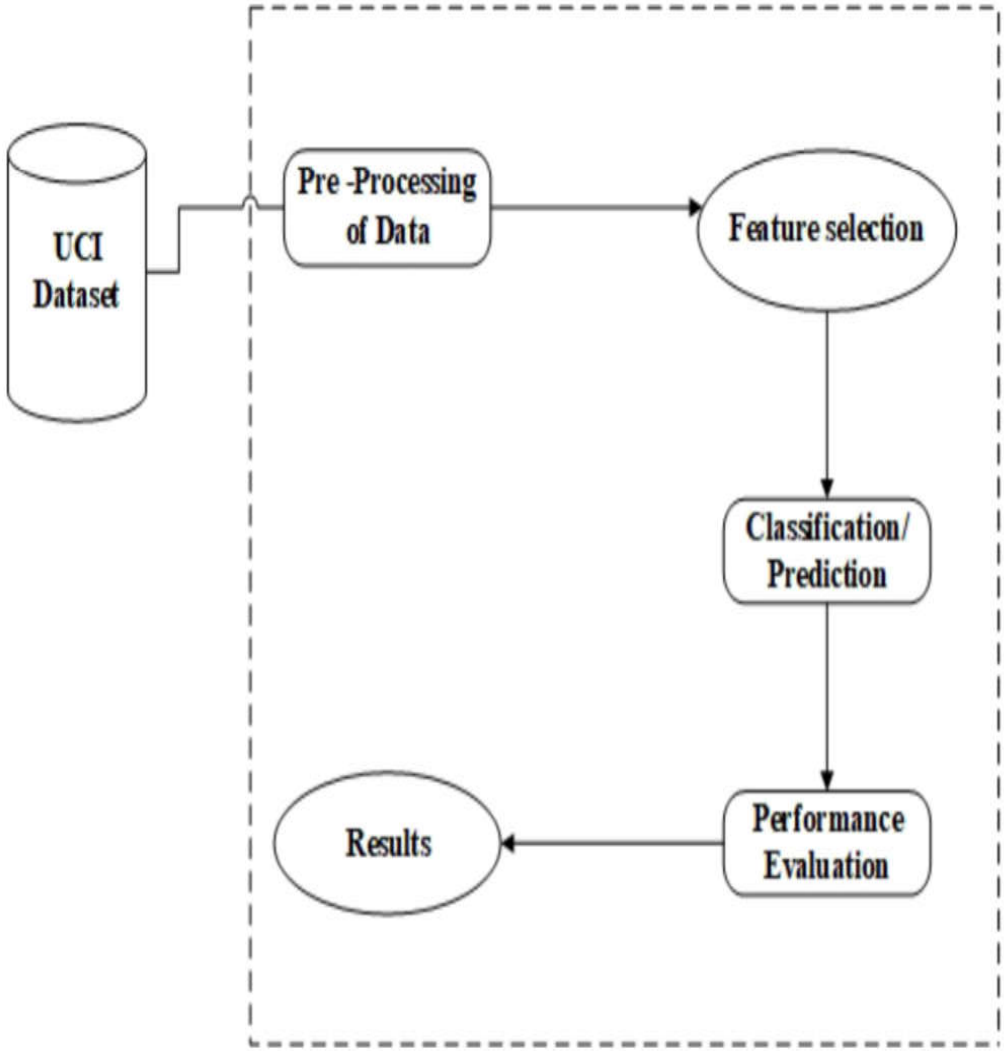
The results are compared for performance and accuracy with these algorithms. The proposed hybrid method returns results of 86.8% for F-measure, competing with the other existing methods.

4. PROPOSED SYSTEM

we propose a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques.

Advantages: We produce an enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM).

5. SYSTEM ARCHITECTURE:



6. IMPLEMENTATION

1) Upload Module: using this module we will upload heart disease dataset of previous patients
2) Pre-process Module: Using this module we will remove all those records which contains missing values. Dataset will be splitted to two parts called training and testing, all classifier will build train model using training data and then test train model by applying test data on that train model to get classification accuracy.
3) SVM Module: Using this module we will build train model using SVM algorithm and then apply test data on that SVM model to get classification accuracy.
4) Naïve Bayes: Using this module we will build train model by using Naïve Bayes algorithm and apply test data to get Naïve Bayes classification accuracy.
5) Logistic Regression: Here train model accuracy will be check with Logistic Regression algorithm
6) ANN Module: Deep Learning Artificial Neural Network train model will be generated and its accuracy can be calculated using test data.
7) HRFLM: Propose Hybrid Algorithm which is combination of Linear model and Random Forest algorithm. Hybrid model will be generated by using both algorithms and then Voting classifier will be used to choose best performing algorithm.
8) Extension Extreme Machine Learning Module: This is an extra module which is built for extension purpose and this module is based on advance Extreme Machine Learning algorithm which can get better prediction accuracy compare to all algorithms. Extreme Learning Machine (ELM) is a novel method for pattern classification as well as function approximation. This method is essentially a single feed forward neural network; its structure consists of a single layer of hidden nodes, where the weights between inputs and hidden nodes are randomly assigned and remain constant during training and predicting phases. On the contrary, the weights that connect hidden nodes to outputs can be trained very fast. Experimental studies in the literature showed that ELMs can produce acceptable predictive performance and their computational cost is much lower than networks trained by the back-propagation algorithm.
9) Graph: This module display accuracy of all algorithms in graph format as comparison

7. Algorithm Details

SVM Algorithm: Machine learning involves predicting and classifying data and to do so we employ various machine learning algorithms according to the dataset. SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyper plane which separates the data into classes. In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification. As a simple example, for a classification task with only two features (like the image above), you can think of a hyper plane as a line that linearly separates and classifies a set of data.

Intuitively, the further from the hyper plane our data points lie, the more confident we are that they have been correctly classified. We therefore want our data points to be as far away from the hyper plane as possible, while still being on the correct side of it.

So when new testing data is added, whatever side of the hyperplane it lands will decide the class that we assign to it.

Random Forest Algorithm: it's an ensemble algorithm which means internally it will use multiple classifier algorithms to build accurate classifier model. Internally this algorithm will use decision tree algorithm to generate it train model for classification.

Decision Tree Algorithm: This algorithm will build training model by arranging all similar records in the same branch of tree and continue till all records arrange in entire tree. The complete tree will be referred as classification train model.

Gradient Boosting Algorithm:Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets, and have recently been used to win many Kaggle data science competitions.

Deep Learning ANN Algorithm: An artificial neuron network (ANN) is a computational model based on the structure and functions of biological neural networks. Information that flows through the network affects the structure of the ANN because a neural network changes - or learns, in a sense - based on that input and output.

ANNs are considered nonlinear statistical data modelling tools where the complex relationships between inputs and outputs are modelled or patterns are found.

ANN is also known as a neural network.

An ANN has several advantages but one of the most recognized of these is the fact that it can actually learn from observing data sets. In this way, ANN is used as a random function approximation tool. These types of tools help estimate the most cost-effective and ideal methods for arriving at solutions while defining computing functions or distributions. ANN takes data samples rather than entire data sets to arrive at solutions, which saves both time and money. ANNs are considered fairly simple mathematical models to enhance existing data analysis technologies.

ANNs have three layers that are interconnected. The first layer consists of input neurons. Those neurons send data on to the second layer, which in turn sends the output neurons to the third layer.

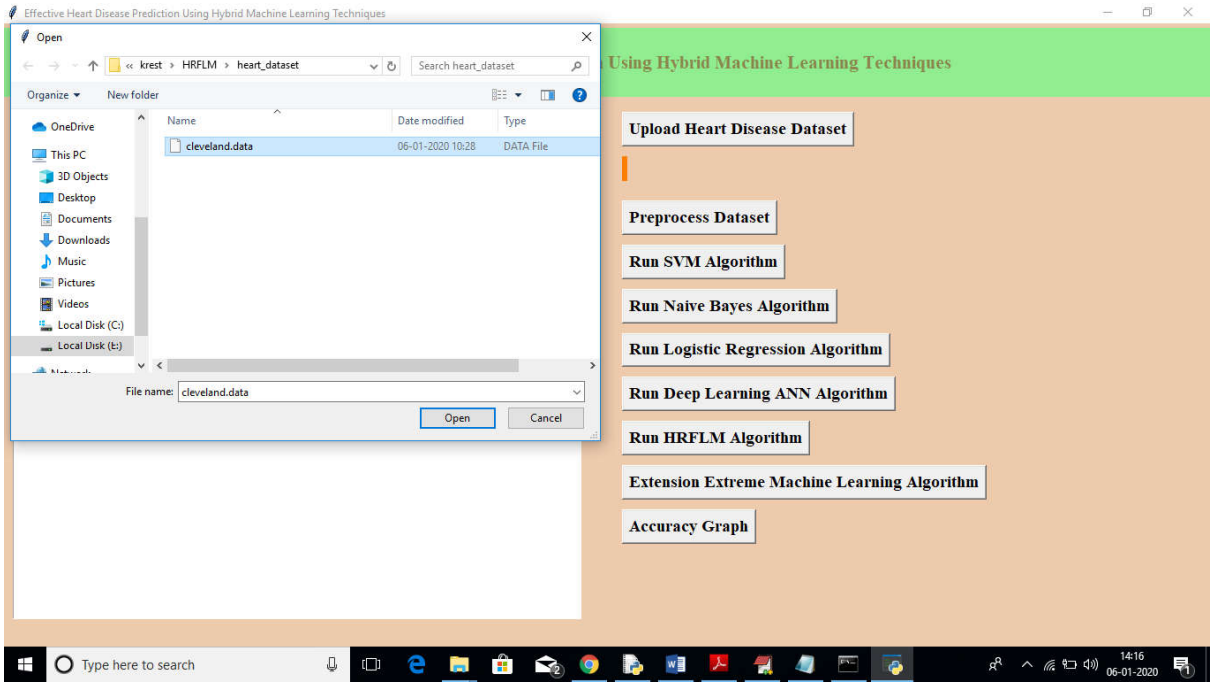
Training an artificial neural network involves choosing from allowed models for which there are several associated algorithms.

All above describe algorithms are to predict heart disease and using this paper we are comparing their performance.

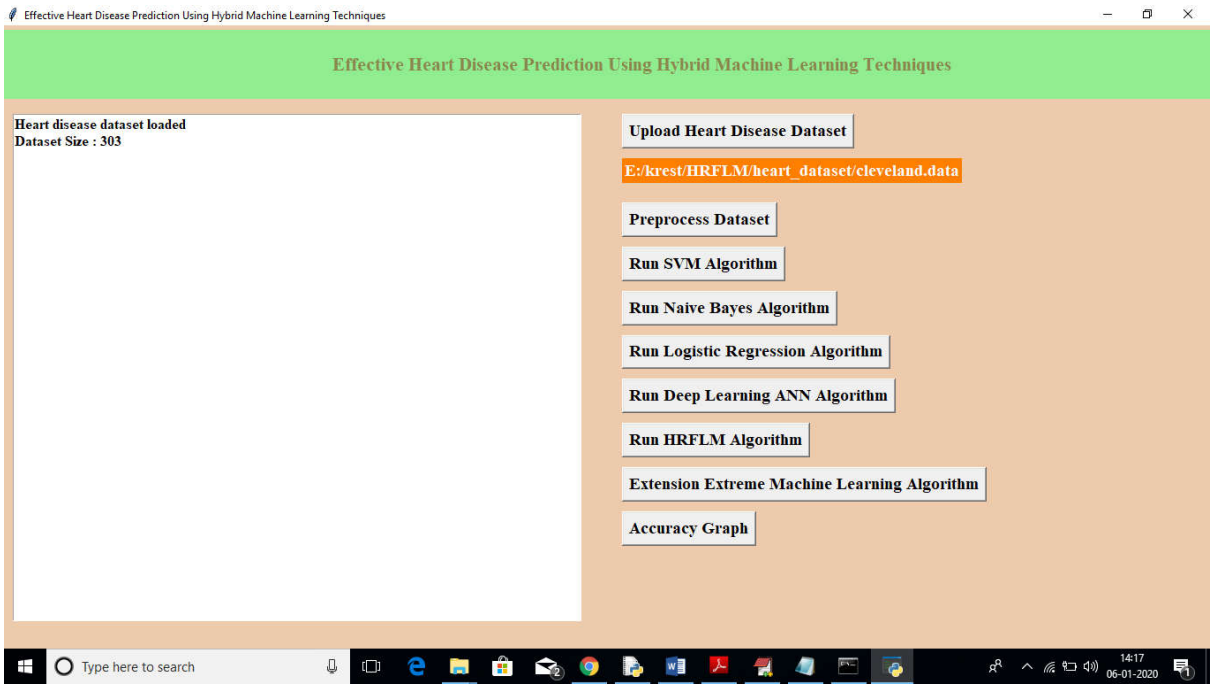
8.SCREENSHOTS

To run this project double click on ‘run.bat’ file to get below screen

In above screen click on ‘Upload Heart Disease Dataset’ button to upload heart dataset



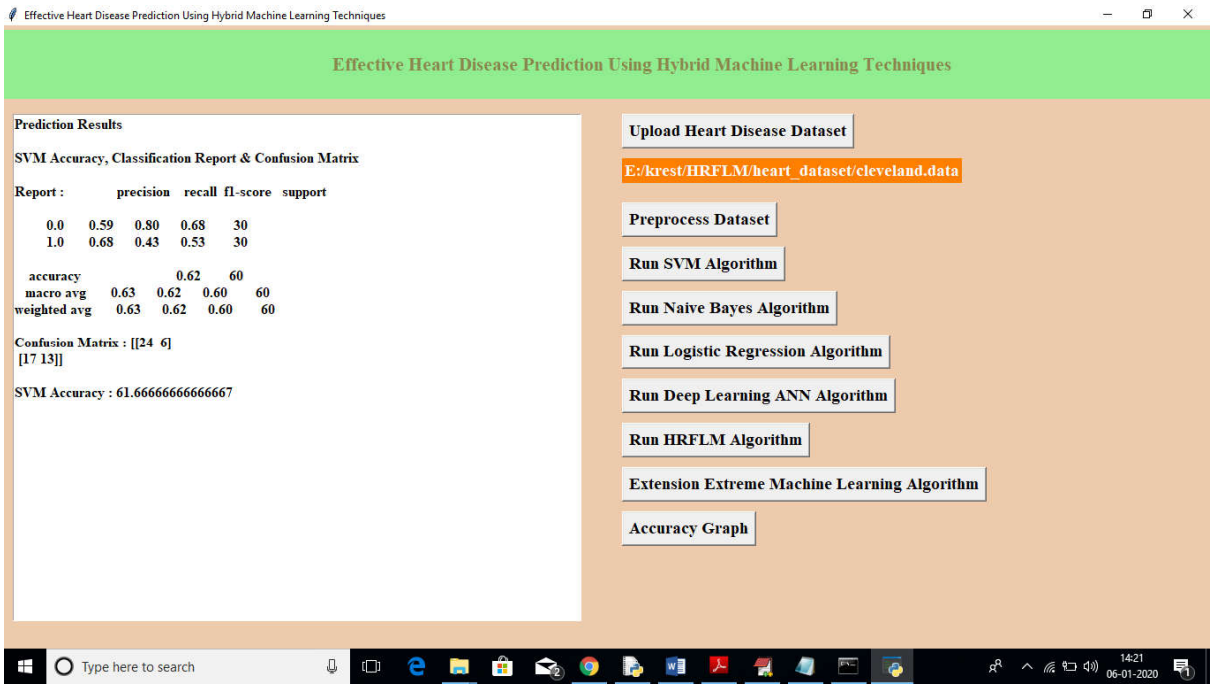
In above screen I am uploading ‘cleveland.data’ dataset, after uploading dataset will get below screen



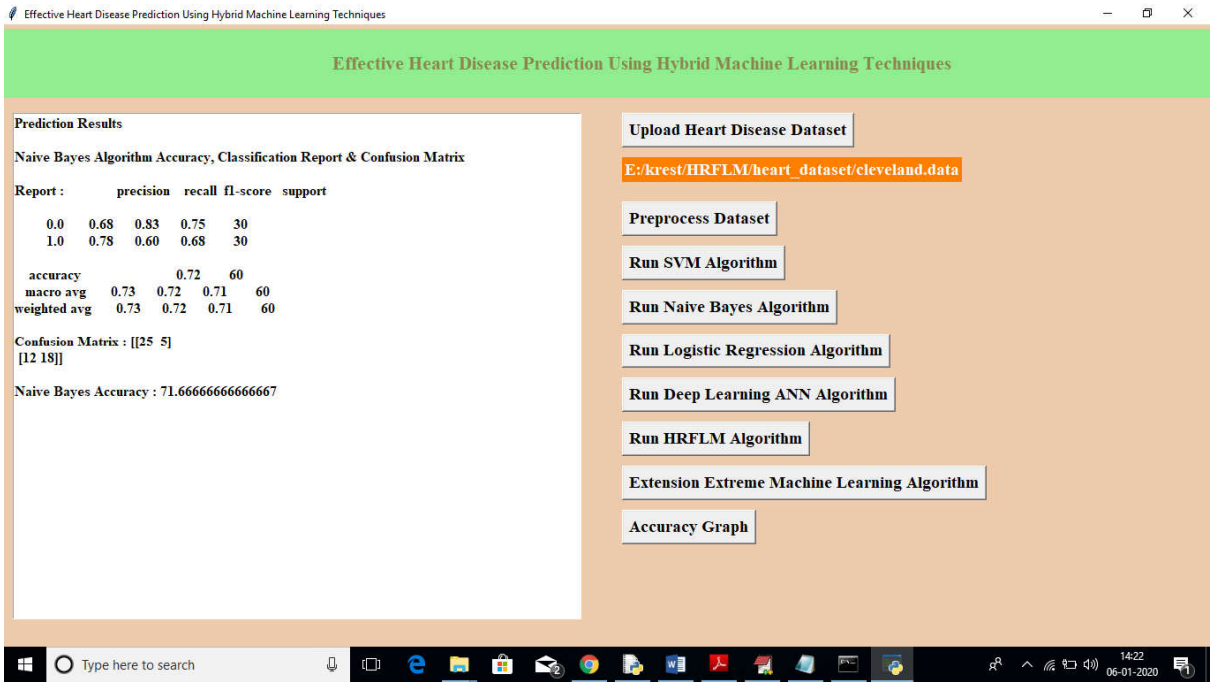
In above screen we can see dataset contains total 303 records, now click on ‘Pre-process Dataset’ button to apply pre-processing technique to remove out all non-numeric data.

In above screen after applying pre-processing dataset size reduced to 297 records and we can see application randomly splitted complete dataset in to tow parts called train and test. For training application using 237 records and for testing application using 60 records. Application will choose random 60 records so always accuracy of same algorithm will be different as records for testing are randomly chooses.

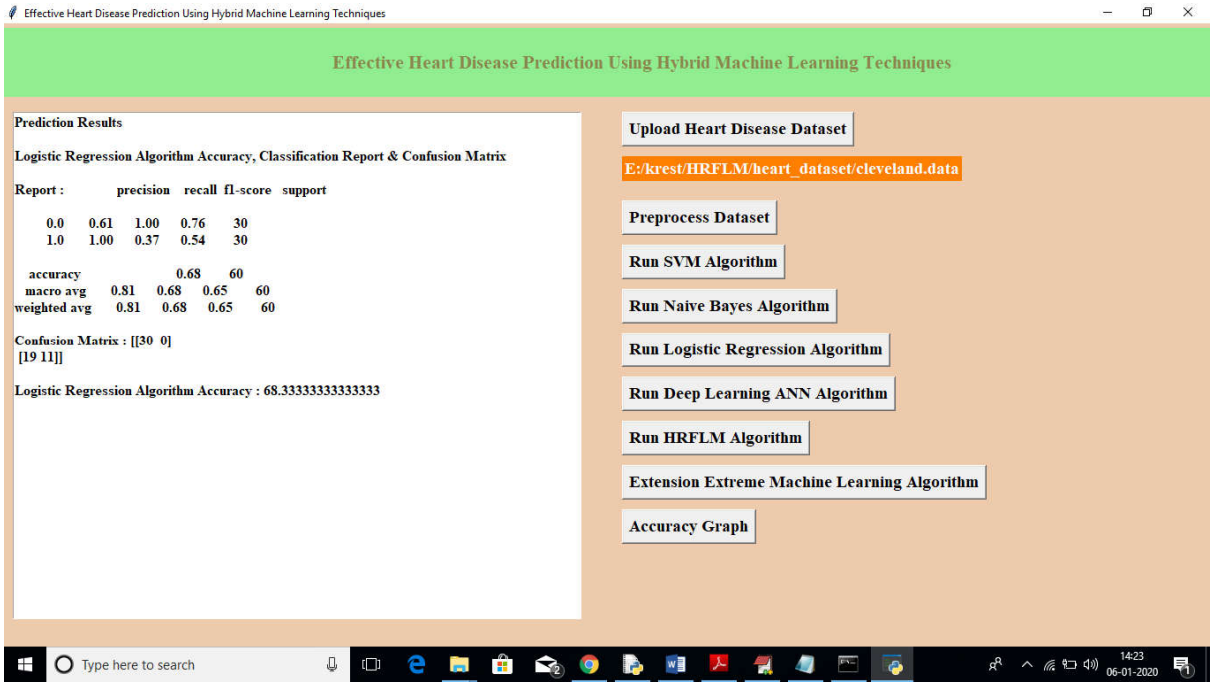
Now click on 'Run SVM Algorithm' button to generate SVM model on train dataset and to apply test data to get SVM classification accuracy.



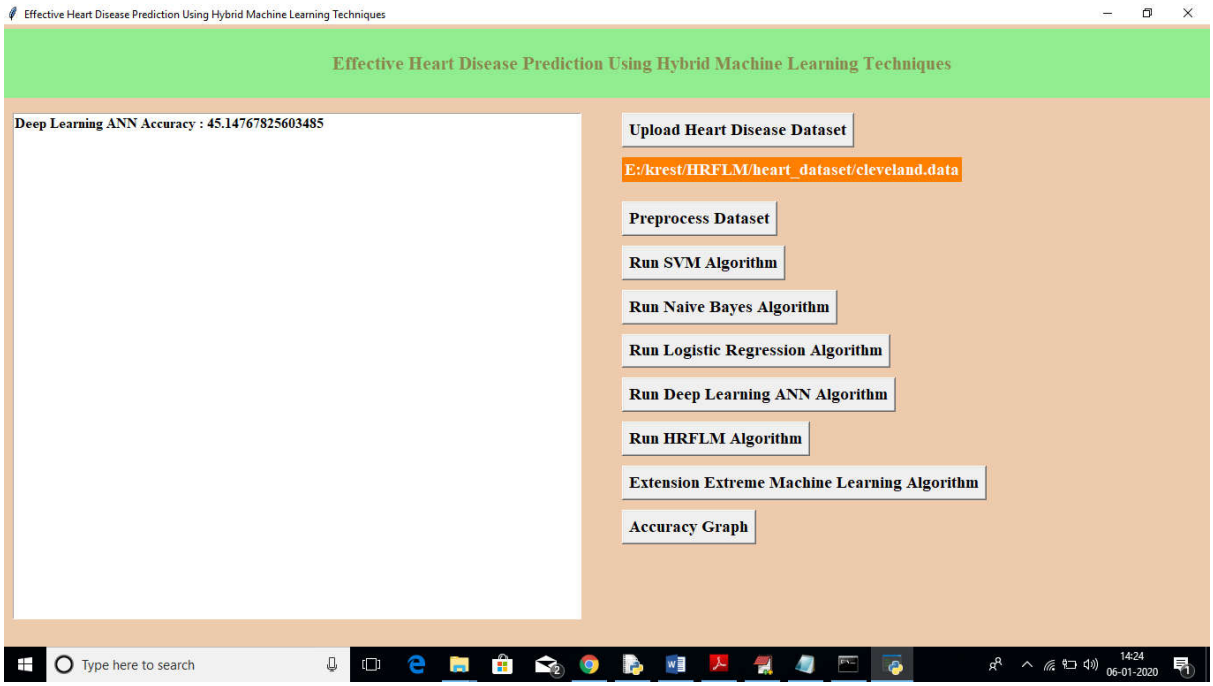
In above screen SVM got 62% accuracy, now click on 'Run Naïve Bayes Algorithm' button to get its accuracy



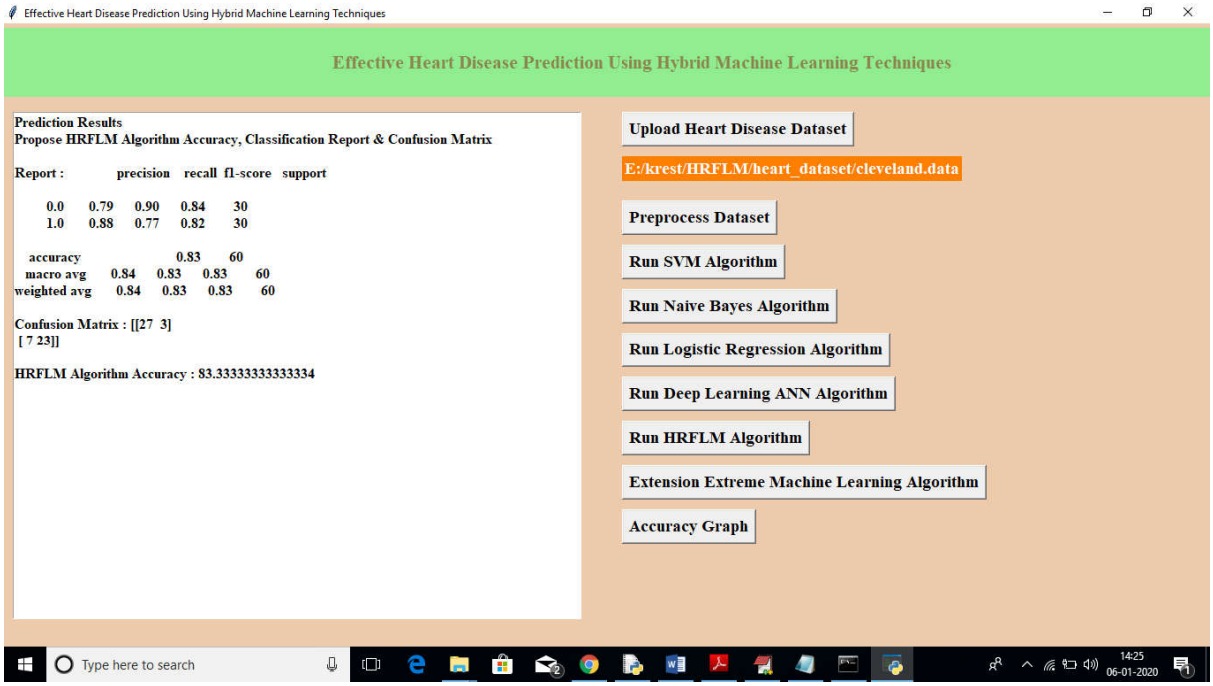
In above screen we can see Naïve Bayes got 72% accuracy, now click on 'Run Logistic Regression Algorithm' to get its accuracy



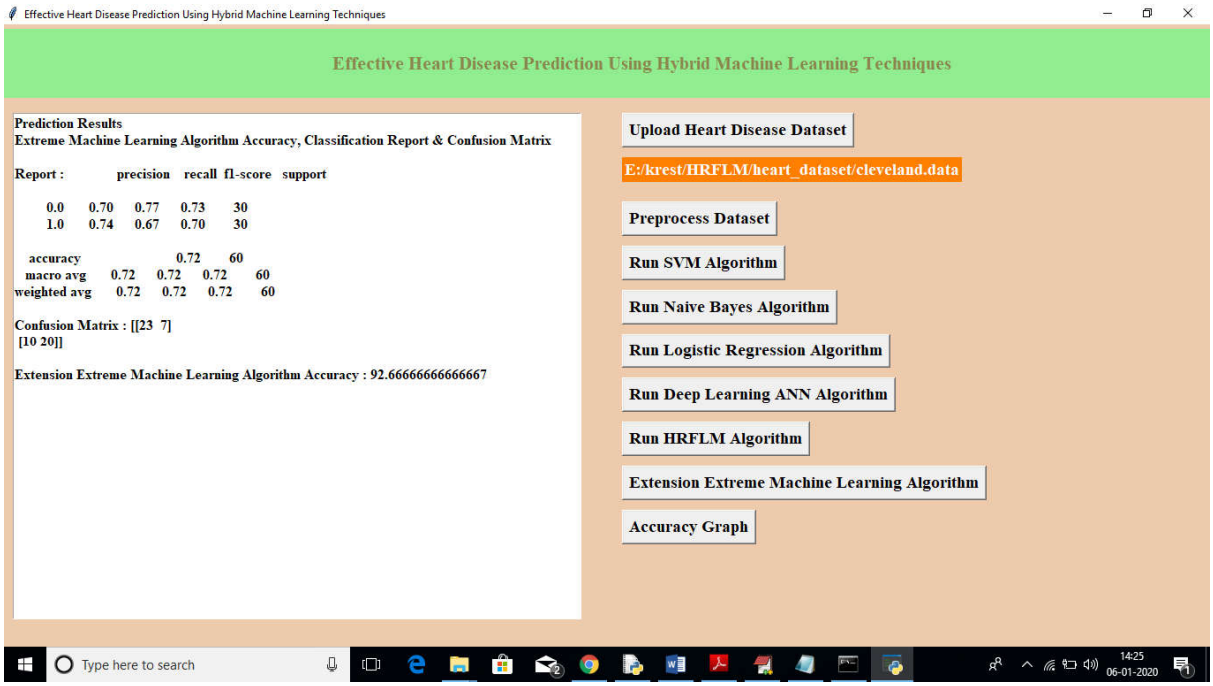
In above screen logistic regression got 69% accuracy, now click on 'Run Deep Learning ANN Algorithm' button to get its accuracy



In above screen we can see ANN got 46% accuracy, now click on 'Run HRFLM Algorithm' button to get propose work accuracy



In above algorithm we can see HRFLM got 84% accuracy, now click on 'Extension Extreme Machine Learning Algorithm' button to check EML extension accuracy



In above screen we can see extension EML algorithm got 93% accuracy which is better than all algorithms. Now click on 'Accuracy Graph' button to get below graph

In above graph x-axis represents algorithm names and y-axis represents accuracy of that algorithm. In all algorithms propose HRFLM and extension algorithm got better accuracy

9.CONCLUSION

Identifying the processing of raw healthcare data of heart information will help in the long term saving of human lives and early detection of abnormalities in heart conditions. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease. Heart disease prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible. Further extension of this study is highly desirable to direct the investigations to real-world datasets instead of just theoretical approaches and simulations. The proposed hybrid HRFLM approach is used combining the characteristics of Random Forest (RF) and Linear Method (LM). HRFLM proved to be quite accurate in the prediction of heart disease.

10. FUTURE ENHANCEMENT

The future course of this research can be performed with diverse mixtures of machine learning techniques to better prediction techniques. Furthermore, new feature selection methods can be developed to get a broader perception of the significant features to increase the performance of heart disease prediction.

11. REFERENCES

[1] A. S. Abdullah and R. R. Rajalaxmi, ‘‘A data mining model for predicting the coronary heart disease using random forest classifier,’’ in Proc. Int. Conf. Recent Trends Comput. Methods, Commun. Controls, Apr. 2012, pp. 22–25.

[2] A. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar, ‘‘Using PSO algorithm for producing best rules in diagnosis of heart disease,’’ in Proc. Int. Conf. Comput. Appl. (ICCA), Sep. 2017, pp. 306– 311.

[3] N. Al-milli, ‘‘Backpropogation neural network for prediction of heart disease,’’ J. Theor. Appl.Inf. Technol., vol. 56, no. 1, pp. 131–135, 2013.

[4] C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika, ‘‘Analysis of neural networks based heart disease prediction system,’’ in Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI), Gdansk, Poland, Jul. 2018, pp. 233–239.

[5] P. K. Anooj, ‘‘Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules,’’ J. King Saud Univ.-Comput. Inf. Sci., vol. 24, no. 1, pp. 27–40, Jan. 2012. doi: 10.1016/j.jksuci.2011.09.002.

[6] L. Baccour, ‘‘Amended fused TOPSIS-VIKOR for classification (ATOVIC) applied to some UCI data sets,’’ Expert Syst. Appl., vol. 99, pp. 115–125, Jun. 2018. doi: 10.1016/j.eswa.2018.01.025.

[7] C.-A. Cheng and H.-W. Chiu, ‘‘An artificial neural network model for the evaluation of carotid artery stenting prognosis using a national-wide database,’’ in Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2017, pp. 2566–2569.

[8] H. A. Esfahani and M. Ghazanfari, ‘‘Cardiovascular disease detection using a new ensemble classifier,’’ in Proc. IEEE 4th Int. Conf. Knowl.- Based Eng. Innov. (KBEI), Dec. 2017, pp. 1011–1014. [9] F. Dammak, L. Baccour, and A. M. Alimi, ‘‘The impact of criterion weights techniques in TOPSIS method of multi-criteria decision making in crisp and intuitionistic fuzzy domains,’’ in Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE), vol. 9, Aug. 2015, pp. 1–8.

[10] R. Das, I. Turkoglu, and A. Sengur, ‘‘Effective diagnosis of heart disease through neural networks ensembles,’’ Expert Syst. Appl., vol. 36, no. 4, pp. 7675–7680, May 2009. doi: 10.1016/j.eswa.2008.09.013.

[11] M. Durairaj and V. Revathi, ‘‘Prediction of heart disease using back propagation MLP algorithm,’’ Int. J. Sci. Technol. Res., vol. 4, no. 8, pp. 235–239, 2015.

[12] M. Gandhi and S. N. Singh, ‘‘Predictions in heart disease using techniques of data mining,’’ in Proc. Int. Conf. Futuristic Trends Comput. Anal. Knowl. Manage. (ABLAZE), Feb. 2015, pp. 520–525.

[13] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, ‘‘Prediction of heart disease using machine learning,’’ in Proc. 2nd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA), Mar. 2018, pp. 1275–1278.

[14] B. S. S. Rathnayakc and G. U. Ganegoda, ‘‘Heart diseases prediction with data mining and neural network techniques,’’ in Proc. 3rd Int. Conf. Converg. Technol. (I2CT), Apr. 2018, pp. 1–6.

[15] N. K. S. Banu and S. Swamy, ‘‘Prediction of heart disease at early stage using data mining and big data analytics: A survey,’’ in Proc. Int. Conf. Elect., Electron., Commun., Comput. Optim. Techn. (ICEECCOT), Dec. 2016, pp. 256–261.

[16]. Shobarani, R., Sharmila, R., Kathiravan, M. N., Pandian, A. A., Chary, C. N., & Vigneshwaran, K. (2023, April). Melanoma Malignancy Prognosis Using Deep Transfer Learning. In 2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1) (pp. 1-6)