

Heart Disease Prediction using Hybrid Machine Learning Techniques

K N V P Ramesh, Asst. Prof CSE: BVCE, Email: ramesh.kb17@gmail.com

P Ramesh, Asst. Prof CSE: BVCE, Email: panthaniramesh.542ram@gmail.com

Goli Manisha, CSE:BVCE

Gonamanda Anupama, CSE:BVCE

Gudimetla Rajendra Sai Nithin, CSE:BVCE

Manepalli Mohan Durga Prasad, CSE:BVCE

Received 2022 March 25; **Revised** 2022 April 28; **Accepted** 2022 May 15.

ABSTRACT:

One of the leading causes of death in the modern world is coronary artery disease. Clinical data analysis faces a major challenge in predicting cardiovascular disease. It has been proven that machine learning (ML) can be used to make predictions and decisions based on the vast amounts of data generated by the healthcare industry. ML techniques have also been used in recent developments in various IoT areas (IoT). Only a sliver of the potential of ML to predict heart disease has been explored so far in various studies. Machine learning techniques are used to find significant features in this paper, which improves the accuracy of cardiovascular disease prediction. A variety of feature combinations and well-known classification techniques are used to build the prediction model. The prediction model for heart disease using the hybrid random forest with a linear model produces an improved performance level with an accuracy level of 88.7 percent (HRFLM).

Keywords: KNN: K-Nearest Neighbor Algorithm; RBFN: Radial Basis Function Network; PSO: Particle Swarm Optimization algorithm; HRFLM: hybrid random forest with linear model.

I. INTRODUCTION:

Heart disease is difficult to diagnose because of a variety of risk factors, including diabetes, high blood pressure, high cholesterol, abnormal pulse rate, and many others. The severity of heart disease in humans has been assessed using a variety of data mining and neural network techniques. These methods include KNN, Decision Trees, Genetic Algorithm and Naive Bayes [11–13] to classify the severity of the disease. [11–13] Because of the complexities of heart disease, it must be treated with great care. If you don't, you run the risk of harming your heart and even dying young. Metabolic syndromes can be discovered using medical science and data mining techniques. Prediction of heart disease and data investigation benefit greatly from data mining with classification.

We've also seen decision trees used to predict the accuracy of heart disease events [1]. For the prediction of heart disease, various methods of data mining have been used to abstract knowledge. Many readings were done to produce a prediction model using a variety of methods, some of which were distinct from one another but which were also linked to one another. Hybrid methods [14] are the name given to these new techniques that combine multiple existing ones. Heart rate time series are used to introduce neural networks. It is possible to use this method to determine the patient's exact condition in relation to heart disease by looking at a variety of clinical records, such as those showing the presence of conditions such as Left bundle branch block (LBBB), Right bundle branch block (RBBB), Atrial fibrillation (AFIB), Normal sinus rhythm (NSR), Sinus bradycardia (SBR), and Atrial flutter (AFL). A radial basis function network (RBFN) is used to classify the dataset, with 70% of the data used for training and the remaining 30% for classification [4, 15].

A heart attack

Heart disease is a medical condition that affects the heart's normal function and function. Narrowing or blockage of a blood vessel can result in a heart attack, angina (chest pain caused by a reduction in blood flow to the heart), or a stroke.

Additionally, there are many other types of heart disease, such as those affecting the muscles and valves of the heart as well as heartbeat rhythm. When it comes to the symptoms of heart disease, it all depends on the type of disease you have. Symptoms of heart disease may differ between sexes. Males are more likely to experience chest pain, while females are more likely to experience symptoms like shortness of breath, nausea, and extreme exhaustion related to their chests. If the blood vessels in your legs or arms are obstructed, you may experience pain, numbness, weakness, or coldness [1]. In addition to these symptoms, pain can be felt in various parts of the body. Arrhythmias, heart defects, weak heart muscles (dilated cardiomyopathy), infections, and vascular heart diseases are all symptoms of heart disease. cardiovascular disease may not be diagnosed until you experience symptoms such as angina or a heart attack. It's important to keep an eye out for signs of heart disease and talk to your doctor about them. Heart disease can be detected early and treated more effectively if discovered early. Any time you experience symptoms such as shortness of breath, dizziness, or fainting, you should go to the hospital right away. Many types of heart disease are preventable or treatable by making healthy lifestyle choices.

II. Predicting the likelihood of an event

Risk forecast instruments are designed to identify patients who are at risk and to motivate doctors to be more proactive in caring for them. This information can be used to select the most appropriate/suggested strategy. Patients' motivation and adherence would be improved if the decision was made as a team effort. A study found that using a risk forecast device prompted doctors to pay closer attention to the results, to learn more about risk factors, and to adopt a more positive attitude toward proactive treatment.

Motivation:

Heart disease prediction models are the primary goal of this study, which is why this research is being conducted.. In addition, the goal of this study is to find the best algorithm for classifying patients who may be at risk of heart disease. Classification algorithms such as Nave Bayes, Decision Tree, and Random Forest are studied and compared at various levels of evaluation to support this work. Despite the fact that these machine learning algorithms are widely used, predicting heart disease is a crucial task that demands the highest level of accuracy possible. It is thus possible to evaluate the three algorithms in a variety of ways. As a result, medical researchers and practitioners will be able to make more informed decisions.

III. Statement of the Issue:

The detection of heart disease is a major challenge. However, while there are instruments that can estimate a person's risk of heart disease, they are either prohibitively expensive or inefficient. It is possible to reduce the risk of death and overall complications by detecting cardiac disease early on. Patients can't be monitored every day, and 24-hour consultations by a doctor aren't possible because they require a lot of time, expertise, and patience. Machine learning algorithms can be used to search for hidden patterns in the massive amounts of data we have today. Hidden patterns in medicinal data can be used for health diagnosis.

The primary objective of this study is to create a model for heart disease that is more accurate and more precise. New patients are quickly identified, diagnostic time is reduced, heart attacks are prevented, and lives are saved.

SVM, Nave Bayes, and Logistic Regression are some of the algorithms we're testing to see how well they classify and predict heart disease. All of these algorithms are capable of making predictions, but their precision falls short of what is required. A new algorithm called Hybrid Machine Learning was developed to improve the accuracy of heart dataset prediction by combining two classification algorithms such as Linear Model and Random Forest. Internally, a voting classifier will be built using Linear Model and Random Forest, and the hybrid algorithm will evaluate the accuracy of prediction provided by each algorithm and give preference to the one with the best accuracy. As a result, we will always have a more accurate algorithm for predicting heart disease if we use a hybrid model.

Using the UCI ML repository's banknote authentication dataset and three different train test ratios (80/20, 60/40, 70/30), this paper applies the SML algorithms of SVM, LR, NB, DT, RF and KNN to the dataset. Attributes in the dataset include 1372 for features and 5 for the target attribute, which has a value as either genuine bank currency or fake currency.

IV. Proposed System

Hybrid Machine Learning is a new algorithm that combines Linear Model and Random Forest classification algorithms to improve the accuracy of heart dataset predictions. It will be possible to create a hybrid algorithm that uses both a linear model and a random forest to build up an internal voting classifier while also evaluating the prediction accuracy of both algorithms and then voting for the one with the best accuracy. As a result, we will always have a more accurate algorithm for predicting heart disease if we use a hybrid model.

4.1 Algorithms:

SVM, Random Forest, Decision Tree, HRFLM, Gradient Boosting, Deep Learning ANN, and Extreme Machine Learning are all used in this study.

4.1.1 SVM Algorithm:

We use various machine learning algorithms depending on the dataset in order to predict and classify data. Classification and regression problems can be solved using the SVM or Support Vector Machine, which is a linear model. Linear and non-linear problems can be solved using this tool. SVM is a simple concept: Using an algorithm, the data is divided into classes by a line or hyper plane. The RBF kernel is a common kernel function in various kernelized learning algorithms in machine learning. Support vector machine classification is one area where it is frequently employed. A hyper plane can be thought of as a line that separates and classifies a set of data in a linear fashion, as shown in the image above for a classification task with only two features.

Algorithm 4.1.2 Random Forest Algorithm

In order to construct an accurate classifier model, this is an ensemble algorithm, which uses multiple classifier algorithms internally. Internally, the decision tree algorithm will be used to generate a classification training model. Algorithm for Making Decisions Through a Decision Tree. By putting all records that are similar in the same branch of the tree, this algorithm builds a training model that can be used to build new models. This process continues until all records are arranged in the entire tree. Classification train model is the name given to the entire tree.

4.1.2. HRFLM

Combining Random Forests and Linear Methods, the proposed hybrid HRFLM approach is applied (Logistic Regression). Heart disease can be accurately predicted using HRFLM. The hybrid technique has three steps:

- i) Identifying the output probabilities of each model. We are using the probability function, which returns the target's probabilities in an array, to implement this. The number of categories in the target variable is the same as the number of probabilities in each row.
- ii) Using the log loss function, determining the optimal weight for combining the two models to achieve the lowest possible classification error. A metric called the "log loss function" measures the degree to which your prediction differs from the actual label.
- iii) Finally, using the weighted average from the previous step, combining the two models, and then predicting.

4.1.3 Iterative Gradient Enhancement Algorithm:

Machine learning algorithms known as gradient boosting classifiers group together a number of previously underperforming learning models in order to create a more accurate and reliable prediction model. Gradient boosting is a common use for decision trees. Gradient boosting models have recently been used to win many Kaggle data science competitions because of their ability to classify complex datasets.

Algorithm for Deep Learning Using ANNs:

ANNs are computer models inspired by biological neural networks in their structure and operation. An ANN's structure can change as a result of the input and output it receives, because a neural network is constantly evolving.

Nonlinear statistical data modelling tools, such as ANNs, are used to model or discover patterns in the complex relationships between inputs and outputs.

An ANN, or neural network, is a type of ANN.

While there are numerous advantages to using an ANN, one of the most well-known is the fact that it can learn from the data sets it is fed. As a result, ANN serves as a tool for approximating random functions. When defining computing functions or distributions, these types of tools help determine the most cost-effective and ideal methods for arriving at solutions. To save both time and money, ANN uses data samples rather than entire data sets to arrive at solutions. As simple mathematical models to enhance existing data analysis techniques, ANNs are widely accepted.

The three layers of an ANN are all connected to each other. The input neurons make up the first layer. The first layer of neurons sends information to the second layer, which in turn sends the output neurons to the third layer.

An artificial neural network is trained by selecting from a variety of models and algorithms.

This paper compares the performance of the algorithms described above in terms of their ability to predict heart disease.

Extreme Machine Learning: An Extension

The Extreme Machine Learning algorithm used in this module is an advanced version of the algorithm used in the original Extreme Machine Learning module. Using the Extreme Learning Machine (ELM), a new approach to pattern recognition and function approximation has been developed. The weights between inputs and hidden nodes are randomly assigned and remain constant throughout the training and prediction phases of this method, which is essentially a single feed forward neural network. Weights that connect hidden nodes to outputs, on the other hand, are able to be trained quickly. Researchers have found that ELMs can produce acceptable predictive performance, and their computational costs are lower than those of networks trained using the back-propagation method.

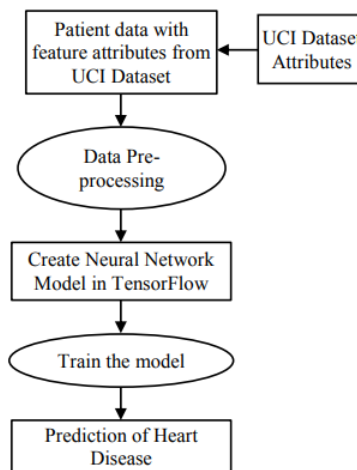


Fig.1 Architecture Diagram

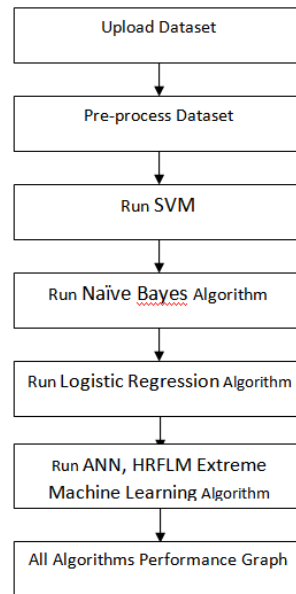


Fig.2. Algorithm and Process a graphical representation of the design concepts.

First, we will use the Upload Module to upload data from previous patients with heart disease.

Using the pre-process module, we can remove all records that have missing data. To determine classification accuracy, the dataset will be divided into two sections referred to as "training" and "testing." Each classifier will build a train model using training data and then test the train model using test data.

In this module, we will build a train model using the SVM algorithm, and then apply test data to that model to see how accurate it is at classifying new data.

When using the Nave Bayes algorithm, we'll build a train model using test data to get the Nave Bayes classification accuracy.

The Logistic Regression algorithm will be used to test the model's accuracy here.

It will be possible to calculate the accuracy of the Deep Learning Artificial Neural Network train model through the use of test data.

Linear and Random Forest algorithms can be combined to create a new hybrid algorithm called HRFLM. In order to select the best performing algorithm, a hybrid model will be built using both algorithms.

In order to extend Extreme Machine Learning, an additional module has been created, which is based on an advanced Extreme Machine Learning algorithm that is capable of predicting more accurately than any other method. Using the Extreme Learning Machine (ELM), a new approach to pattern recognition and function approximation has been developed. The weights between inputs and hidden nodes are randomly assigned and remain constant throughout the training and prediction phases of this method, which is essentially a single feed forward neural network. Weights that connect hidden nodes to outputs, on the other hand, are able to be trained quickly. Researchers have found that ELMs can produce acceptable predictive performance, and their computational costs are lower than those of networks trained using the back-propagation method.

9) Graph: As a comparison, this module shows the accuracy of all algorithms as a graph.

Incorporation and Result

4.2 The collection of data

Most researchers use the Cleveland heart disease dataset, which can be found in the UCI online repository for machine learning purposes. There are 303 samples in total, 6 of which are missing values. Originally, the data had 76 distinct features, but the published work almost certainly only mentions 13 of those, with the remaining feature focusing on disease effects.

Table 1.

S. No.	Attribute	Description	Range
1	Age	Age of the individual	29-77
2	Sex	Sex	M, F
3	CP	Chest Pain type	1-typical angina
			2-atypical angina
			3-Non-Anginal Pain
			4-Asymptomatic
4	restbp	Resting Blood Pressure	94-200
5	serchol	Serum Cholesterol in mg/dl	126-564
6	fbs	Fasting blood sugar > 120	Yes, No
7	restecg	Resting Electrocardiographic	0, 1, 2
8	mhr	Maximum Heart rate achieved	71-202
9	exang	Exercise Induced Angina	Yes, No
10	oldpeak	ST depression Induced by Exercise relative to Rest	0-6.2
11	slope	Slope of the Peak Exercise ST Segment	1, 2, 3
12	vca	Number of Major Vessels colored by Fluoroscopy	0, 1, 2, 3

4.2.1 Metrics for Evaluation:

In terms of F1-Score, Accuracy, and Recipient, Areas in which the device can be used We use ROC-AUC metrics to gauge how well our models are performing. FPR=False Positive Rate must be used to evaluate the F1-score, accuracy, precision, and recall.

To put it another way, TPR stands for True Positive Rate.

F1-score: Accuracy, Precision, Recall

Values are calculated in this manner, and the results are evaluated in terms of:

The number of events that constitute a true positive (TP) is the number of events that have been accurately counted.

Unneeded or incorrectly predicted events are known as false negatives (FNs).

Incorrectly predicted number of events is known as a false-positive (FP).

No. of events that were foreseen but not required. (TN)

Machine learning accuracy can be evaluated using the False Positive Rate (FPR), which measures the number of false positives in the system. The formula for calculating the FPR is: $FP/(FP+TN)$

a measure of the accuracy of a test result It is a synonym for recall and is therefore defined as $TPR=FP/(FP+TN)$.

Simply dividing the number of correctly predicted observations by the total number of observations is an easy way to measure accuracy.

$Accuracy=(TN+TP)/(TP+FP+TN+FN)$

It's the ratio that accurately predicts positive observations in the original data.

$TP/(TP+FN) = Recall$

In order to get the most accurate results, precision is needed. In other words, this means determining the total number of software's predicted to be positive that are actually positive. There are two ways to look at precision: through the prism of time and space.

F1-score: The F-score is a way of combining precision and recall in a machine learning model. high levels of accuracy and recall Precision and recall are defined as the <https://deeptai.org/machine-learning-glossary-and-terms/harmonic-mean> of the model, and it is Precision and recall are two important properties of the model, and they are both described by the term "harmonic mean" at <https://deeptai.org>. The F-score is another name for it. Precision Recall/Precision + Recall is the formula used to calculate F1 Score.

If you're trying to solve a classification problem, you'll want to keep an eye out for metrics that can help you determine how well you're doing.

4.3 Outcome:

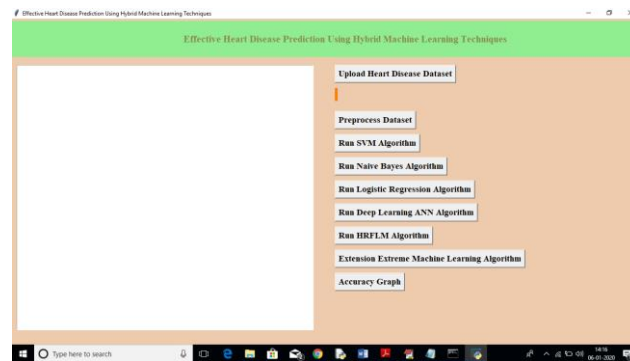


Fig. 3. User Interface

In above diagram click on 'Upload Heart Disease Dataset' button to upload heart dataset

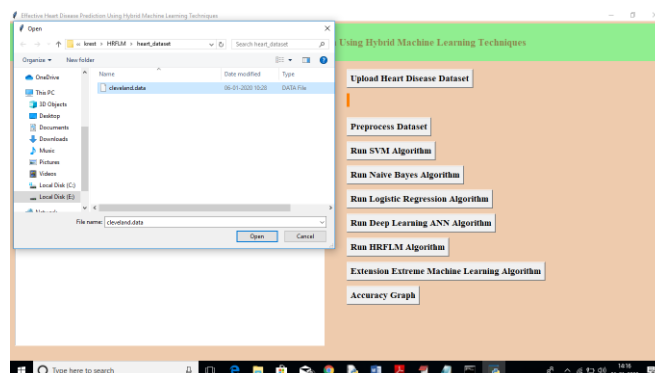


Fig. 4. Uploaded cleveland.data

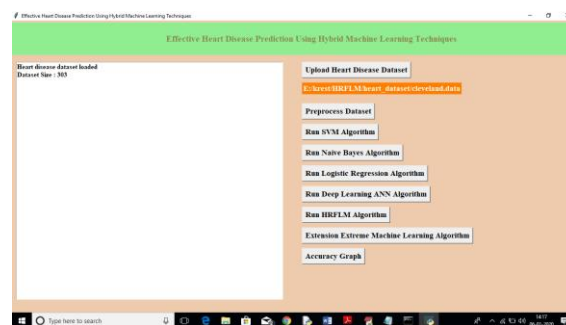


Fig. 5 Preprocessing Data

In above diagram 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.

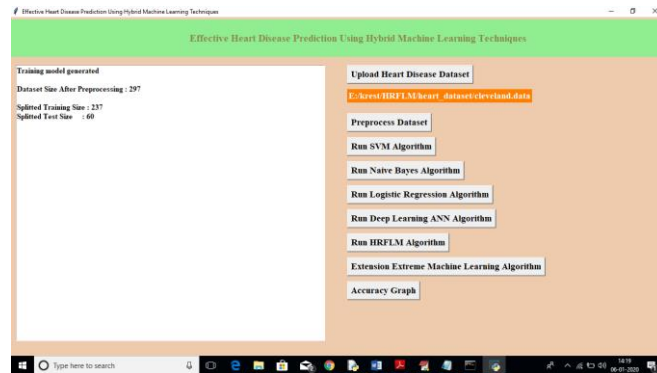


Fig. 6. Size Reduction

In above diagram 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.

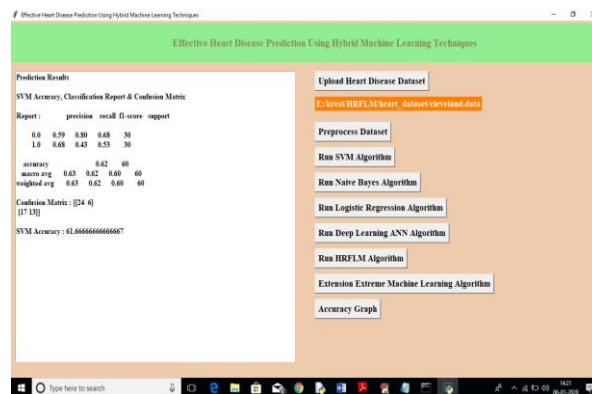


Fig. 7. Run SVM Algorithm

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

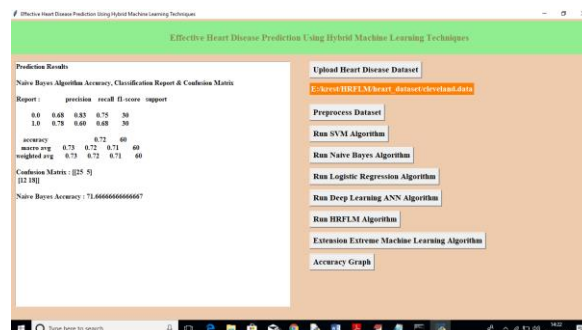


Fig. 8 . Applied Naïve Bayes Algorithm

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

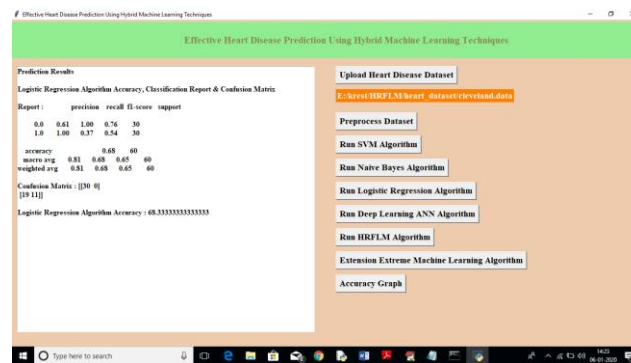


Fig.9. Regression

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

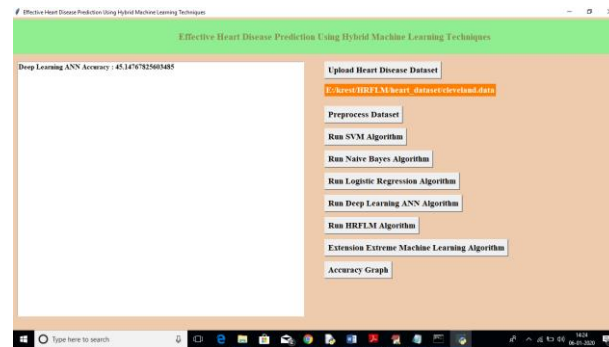


Fig.10 ANN

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

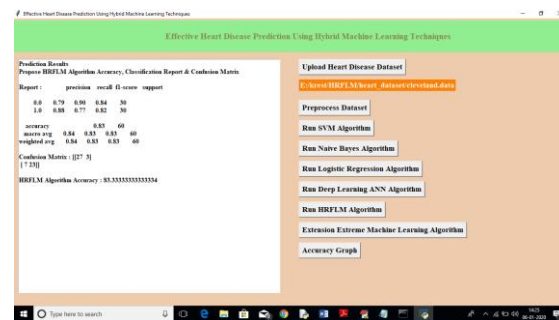


Fig.11 HRFLM Accuracy

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

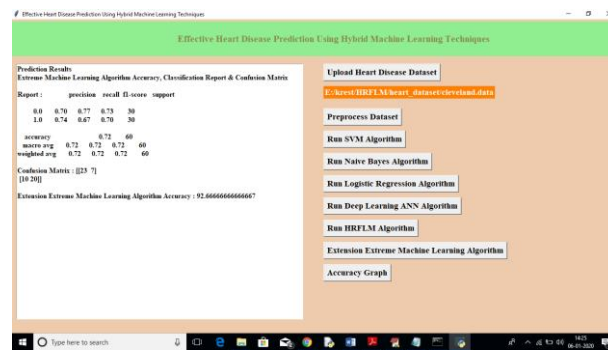


Fig.12 extension EML algorithm

In above diagram 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



Fig.13 Accuracy

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

CONCLUSION

To save lives and detect abnormalities in heart conditions in the long term, it is important to identify how raw healthcare data of heart information is processed. Using machine learning techniques, raw data was processed and a new and novel understanding of heart disease was gained. Prediction of heart disease in the medical field is a difficult and important task. This can be greatly reduced if the disease is detected early and preventative measures are taken as soon as possible. It would be ideal if this research could be expanded in order to focus on real-world datasets rather than merely theoretical ones and computer simulations. The hybrid HRFLM approach combines Random Forest (RF) and Linear Method (LM) characteristics (LM). Predicting heart disease with HRFLM, an Extreme Learning Machine (ELM), was found to be very accurate. Machine learning techniques can be used in a variety of ways in the future to improve prediction methods..

Bibliography

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, "Backpropagation 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. Rathnayake and G. U. Ganegoda, "Heart diseases prediction with data mining and neural network techniques," in Proc. 3rd Int. Conf. Convergent 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. (ICEECOT), Dec. 2016, pp. 256–261.
16. J. P. Kelwade and S. S. Salankar, "Radial basis function neural network for prediction of cardiac arrhythmias based on heart rate time series," in Proc. IEEE 1st Int. Conf. Control, Meas. Instrum. (CMI), Jan. 2016, pp. 454–458.
17. V. Krishnaiah, G. Narsimha, and N. Subhash, "Heart disease prediction system using data mining techniques and intelligent fuzzy approach: A review," Int. J. Comput. Appl., vol. 136, no. 2, pp. 43–51, 2016.
18. P. S. Kumar, D. Anand, V. U. Kumar, D. Bhattacharyya, and T.-H. Kim, "A computational intelligence method for effective diagnosis of heart disease using genetic algorithm," Int. J. Bio-Sci. Bio-Technol., vol. 8, no. 2, pp. 363–372, 2016.
19. M. J. Liberatore and R. L. Nydick, "The analytic hierarchy process in medical and health care decision making: A literature review," Eur. J. Oper. Res., vol. 189, no. 1, pp. 194–207, 2008.
20. T. Mahboob, R. Irfan, and B. Ghaffar, "Evaluating ensemble prediction of coronary heart disease using receiver operating characteristics," in Proc. Internet Technol. Appl. (ITA), Sep. 2017, pp. 110–115.