Speculating the Threat of Cardiovascular Disease Using Classifiers with User-Focused Security Evaluations

S.R. Chandrasekaran
Department of Computer Science & Engineering, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, Tamil Nadu, India.
E-mail: sekaranchandru2000@gmail.com

Dr.N. Sabiyath Fatima
Associate Professor, Department of Computer Science & Engineering, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, Tamil Nadu.
E-mail: sabiyathfathima@crescent.education

Received October 07, 2021; Accepted December 24, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V19I1/WEB19372

Abstract

In recent decades, cardiovascular disease (CVD) is the most common type of disease that is prevailing all over the world. It is a class of diseases that involve the heart and its vessels. Strokes and heart attacks are normally critical events that are largely provoked by congestion that restricts blood from streaming to the parts of the body. The principle aim of this research is to find the feature that accounts for cardiovascular disease risks. The collection of data from the hospitals and laboratories can determine the risk of patients having cardiovascular disease by analysing the trends and correlations between the dataset. First, the data undergoes a security process that involves user-level security. The data is further processed to show the comparison between the 12 features and find out the top few features that account for the risk of positive cardiovascular disease. This Machine learning techniques can be widely used in the medical field, to determine the risk of cardiovascular disease early. These collected data from the patients may be used to warn them. To identify the risk of having cardiovascular disease early there are some proposed machine learning algorithms like K-nearest neighbours, XG Boost, Gradient boost and Random Forest Classifier by measuring the metrics like precision, accuracy, f-1 score and recall. Out of all these algorithms, XG Boost yields the highest accuracy of 90%. To increase the overall accuracy stacking algorithm is used to combine all the base learner algorithms to produce a result at once. After stacking the overall accuracy boosted to 92% for the respective dataset.

Keywords

K-nearest Neighbours, XG Boost, Random Forest, Gradient Boost, Stacking Classifier, Cardiovascular Disease Prediction.

Introduction

Among all other diseases, cardiovascular disease contributes to the world's leading cause of death. Around 17.9 million people died from CVDs in 2016, with global mortality of 32% (WHO). In India during 1990 mortality from CVDs was 2.2 million. Further, it steeply raises to 4.77 million at the end of 2020 which was quite a large number. It constitutes a significant cause called disability. There are a lot of factors that assist heart disease. They are called risk factors. Some of these factors are uncontrollable they are Age, sex, race or ethnicity, history of the family, but there are a few that you can control. Discovering about them can decrease your risk of cardiovascular disease. The hazard determinants such as diabetes, smoking, high cholesterol, Obesity are associated with an increased predominance of coronary heart disease. The suggested work strives to identify these at an immature stage to prevent miserable events. A huge amount of data is available on the internet which was extracted from real people and analyzed by medical professionals. Data mining is a valuable extraction of hidden data from the dataset. In the medical field, mostly the database consists of distinct information. Several machine learning methods can be used for the detection, analysis and prognosis of multiple complaints. The principal intention of this document is to present a system for medical professions to discover the risk of cardiovascular disease at primary stages. Machine learning performs an important role in the discovery of hidden discrete patterns. Regarding the patient's data security, User-level security, a user-focused transmission of data that is added to this paper, thereby interpreting the data, machine learning help in cardiovascular disease prediction and early diagnosis. This paper presents an analysis of several machine learning techniques such as Gradient boost, XGBoost, K-nearest neighbor and Random Forest. Combining all the algorithms with the method called 'Stacking' where it analyzes the data and predicts the overall accuracy from the dataset.

Related Works

Sulabha S. et al., developed an enhanced study of heart disease prediction methods along with the use of data mining classification techniques. They worked on Neural Networks, Decision Trees, and Naive Bayes individually. Their review explains, Neural Networks predicts the disease with the greatest efficiency, but have used only three classification models to predict the accuracy.

Niti Guru et al., designed a system for the prediction of cardiovascular disease, Pressure and glucose level of blood with the assistance of neural networks. Totally 13 attributes were used and the suggested approach calculates the gradient of the error function and is further

used for the data. The drawback was that they have used only 78 patients' data to produce accuracy.

Dundigalla Ravi et al., did a performance analysis that predicts heart disease by exploring four different classification algorithms. The key focus of this method is to efficiently foretell the patients who experience cardiovascular disease. But the shortcoming is that they have taken 14 attributes to predict and have clung to bagging algorithms.

Fahd Saleh Alotaibi et al., has devised a prototype analyzing different ML algorithm. He also used a tool called Rapid miner which gives out immense accuracy contrasted to other tools such as Weka and MATLAB. The highest accuracy was found in the decision tree among other ML algorithms.

Ramya G Franklin et al., submitted a method that controls Navies Bayesian to achieve Smart Heart Disease Prediction to prognosticate the risk determinants concerning heart attack. Additionally, to defend the data transfer the Advanced Encryption Standard algorithm is used for prediction. The main drawback was the usage of only one classification algorithm.

J. Thomas et al., reviewed multiple classification algorithms for diagnosing cardiovascular disease. The systems were Decision tree, Naive Bayes, KNN, Neural network. The final efficiency of the algorithms was examined by interpreting with a different number of attributes. The liability is that usage attributes that are not crucial.

Bum Ju Lee et al., employed a classification system for extracting different features by evaluating the Variability of heart rate from the ECG, heart disease patterns, pre-processing of the given data. The dataset was divided into a couple of groups, namely healthy patients and patients with the positive disease, which were used to inquire with the associative classification algorithms. The paper lacks multi-parametric indices analysing the disease.

Yang Q et al., researched cardiovascular disease based on past family history by using a technique called FHPHD. But the drawback the data were taken was old and outdated.

Solanki et al., studied the perspective of several distribution systems and evaluated them, particularly Naïve Bayes, SVM, Decision tree, K-nearest neighbor, and indeed hybrid approach of the classification algorithms. A review of diverse models confirmed that systems based on classification achieve elevated accuracy as opposed to others. The main weakness is that the model was overfitting.

M.O. Shafiq et al., investigated the outcomes explained that the system performance and effectiveness are adequate with cardiovascular disease analysis. The Random Forest Classifier obtains an accuracy of 92.44% while Naïve Bayes Classifier only got 61.96%, and 59.7% for Logistic Regression. The researchers have only worked on the bagging algorithm.

Tahira Mahboob et al., have submitted several machine learning systems in discovering and observing many cardiovascular diseases. Some ML strategies discussed are Feature Selection, Hidden Markov Models, genetic algorithm, Support Vector Machine, prediction system, data mining methods. This proposed an ensemble ML model employing appropriate techniques which flawlessly classifies various cardiovascular conditions.

Prabhat Pandey et al., have examined the execution of the classification rule algorithms i.e. PART based on K-Means algorithms. Projective Adaptive Resonance Theory is analyzed on the clustered dataset. To estimate the impartial evaluation of the forecast, design a 10-fold cross-validation method was interpreted. Accuracy was comparatively lower when compared to the recent algorithms.

Usman Qamar et al., have suggested an ensemble framework based on a preponderance voting scheme that connects unique classifiers. It also produces more prominent accuracy for heart disease examination. The proposed structure is based on a novel sequence of complex classifiers i.e. Naive Bayes, decision tree-GI and Support Vector Machine.

R.Delshi et al., introduced an algorithm that is used for penetrating topmost periodic models in a heart attack (MAFIA). The periodic designs can be examined by using the statistical classifier (C4.5) algorithm as training using the idea of information entropy. The devised prediction method can predict heart seizures with sufficient effectiveness. The major drawback was that they stuck to the K-means algorithm and developed only within.

Palaniappan S et al., has developed a model using data mining techniques, specifically, Neural Network, Decision Trees, Naïve Bayes called Intelligent Heart Disease Prediction System. Outcomes explain every technique becomes its sole intensity in achieving the purposes of the assigned goals.

Proposed Methodology

The aimed work predicts the risk of cardiovascular disease by exploring the Machine learning classification algorithms and integrating all the analyzed algorithms into a stacking algorithm. Since many studies and published papers had done the same. In this paper, tuned parameters and stacking of all the algorithms were incurred. This research intends to

prognosticate patient who has a risk of cardiovascular ailment. The Doctor, lab technician or nurse collects the data from the patient's health report. Further, the data undergoes a user-level security process that confirms the identity of the person, who requires the data to proceed. Then, it is directed to an ML model, which prophesies the odds of having heart disease. Fig. 1. Displays the working machine learning model for XGBoost, Gradient boost, Random Forest, K-nearest neighbor.

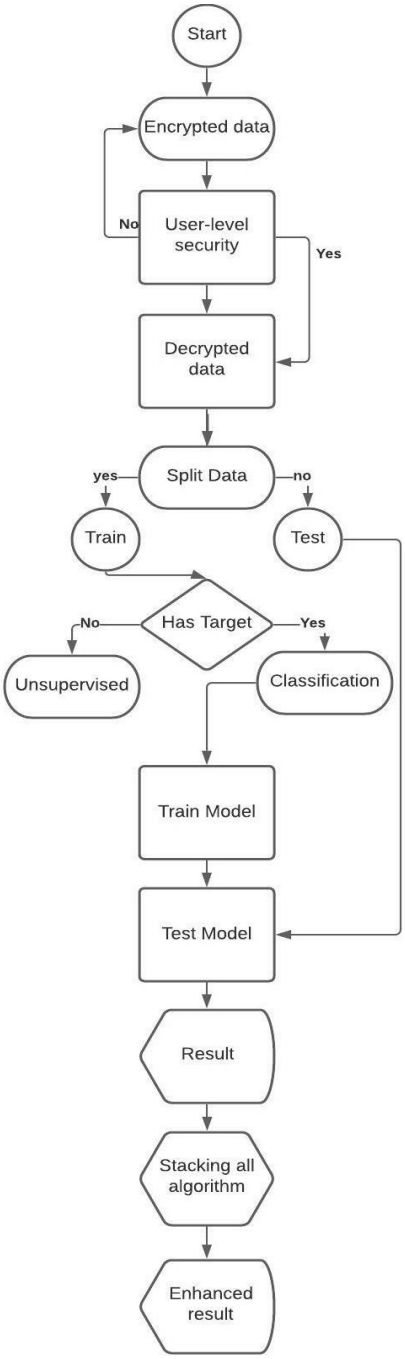

Fig. 1 ML model for prediction

1) Data Extraction

The dataset used in this paper is taken from the website IEEE Dataport, which contains various medical data of the people. The data was collected from different volunteers and tabulated. The dataset mainly consists of 12 features. Although, few published papers used more than 12 features. To determine the risk of CVDs the features that are used in this paper is enough. The description of the features shown in Fig 2.

S.NO.	Features	Format
1	Age	Integer (#)
2	Sex	Male=1; Female=0 (binary)
3	Chest Pain	4 Values (Ordinal)
4	Resting blood pressure	Integer (#)
5	Serum Cholesterol	Integer (#)
6	Fasting Blood Sugar	(> 120 mg/dl) True=1; False=0 (binary)
7	Peak Exercise ST segment	3 Values (Ordinal)
8	Maximum heart rate achieved	Integer (#)
9	Resting Electrocardiographic result	3 Values (0, 1, 2)
10	Exercise induced at rest	Integer (#)
11	Exercise induced angina	Yes=1; No=0 (binary)
12	Target(predictor)	Pos=0, Neg=1

Fig. 2 Features and predictors

2) User-Level Security

The first phase of the dataset begins with the user-focused security evaluation. In this process, the program first checks the username of the person who requested the database. If the username is correct, then it asks for the password of the user. Finally, it processes data regarding the level of user. For instance, if the user is a doctor, the application allows accessing all the data of the patient, but if the user is a nurse or manager the data are limited due to security reasons. Fig 3. shows the data that is only accessible to the doctors, which is all the data are available only for the doctor level.

```
Clinical Data of Patients!

Username: doctor
Password: doc
Doctor Access Granted!!!

       age  sex  cp  trestbps  ...  target        name  patient_id        phone
0       40    1   2       140  ...       0       Jacob       ji78c  418-5291-03
1       49    0   3       160  ...       1       Rosie       jpivt  417-8530-78
2       37    1   2       130  ...       0    Isabella       slopx  360-8955-54
3       48    0   4       138  ...       1    Samantha       yrzk8  716-5884-99
4       54    1   3       150  ...       0  Maximilian       o3a4w  096-7060-05
...    ...  ...  ..       ...  ...     ...         ...         ...          ...
1185    45    1   1       110  ...       1         NaN         NaN          NaN
1186    68    1   4       144  ...       1         NaN         NaN          NaN
1187    57    1   4       130  ...       1         NaN         NaN          NaN
1188    57    0   2       130  ...       1         NaN         NaN          NaN
1189    38    1   3       138  ...       0         NaN         NaN          NaN

[1190 rows x 15 columns]
```

Fig. 3 User-level data for doctor

Fig 4. Shows only the data of patient name, I-D and phone number and restricts all other data showcased in the doctor and other levels of data.

```
Clinical Data of Patients!

Username: manager
Password: man
 Manager Access Granted!!!

             name patient_id        phone
0           Jacob       ji78c  418-5291-03
1           Rosie       jpivt  417-8530-78
2        Isabella       slopx  360-8955-54
3        Samantha       yrzk8  716-5884-99
4     Maximilian       o3a4w  096-7060-05
...          ...         ...          ...
1185         NaN         NaN          NaN
1186         NaN         NaN          NaN
1187         NaN         NaN          NaN
1188         NaN         NaN          NaN
1189         NaN         NaN          NaN

[1190 rows x 3 columns]
```

Fig. 4 User-level data for Manager

Also, to protect the data from unauthorized intrusion the threshold for wrong credentials was set to three. Fig 5. Shows a failure message when the attempts to access the data exceeds three times (max attempts).

```
Clinical Data of Patients!

Username: uvcs
Password: sc
Your login credentials is wrong!


Username: scsc
Password: scs
Your login credentials is wrong!


Username: scss
Password: grg

Attempts exceeded!!!!

Please try again later
```

Fig. 5 Failure attempts

Implementation

1) Data Cleaning and Implementing Machine Learning Models

The data taken from the IEEE data-ports is now subjected into the data cleaning process. Then the data is segregated into some visualization techniques to find the correlation between the data. The proposed process is follows,

Step1: Initially, the data is checked for the null value and the data is compared with each other.
Step2: After comparing the data, a positive relationship is found between chest pain & target i.e., predictor. Since the more elevated measure of pain in the chest ends in a higher risk of having a heart condition.
Step3: Chest pain is an ordinal trait with four chief values: 1st: standard angina, 2nd: abnormal angina, 3rd: non-anginal pain, 4th: symptomless.
Step4: Then, the dataset is compared for positive and negative patients by assessing the top 4 features. In the proposed method, there are huge differences in averages for our features.
Step5: After checking, positive patients encounter an increased value in thalach. However, positive inmates display about one part of three equal parts amount of the oldpeak value.
Step6: Comparing both the symptomatic and asymptomatic patient values and estimating the value for further predictive analysis.
Step7: In creating the Machine learning model, First the data is splitted into two (Test and Train). Then it is further directed to XGBoost, Gradient boost, Random Forest and K-nearest neighbor classifier models.
Step8: From each Machine learning model the final result of trained model is obtained in the form of classification report and confusion matrix.
Step9: To enhance the accuracy of the model stacking algorithm was used, in which all the model undergoes a process and a meta classifier is added to produce the result.

2) Data Preprocessing

The collected data is subjected to the data analysis process, which includes encountering the correlation between the features and the 'target', which is the predictor, in addition to that filtering the positive and negative patients. The inference from the Fig 6. shows, the peak exercise ST segment has a higher value correlated to the target. The peak exercise ST segment has three ordinal values 1 for upsloping and 2 for down sloping. But this is expected in both positive and negative risk patients. The ST segment was nothing but a break between ventricular depolarization and repolarization.

Fig. 6 Correlation matrix

The main usage of the violin plot was to make a better understanding between the positive and negative patients. It can also depict the interquartile range, median, and even outliers of the data. In Fig 7. class=0 was the patients with positive risk displayed a lesser median for ST depression level but at the same time a large data scattered from 0 to 2. Comparatively, class=1 is the patients with negative risk, and data was scattered over 1 and 3. No vast transition between male and female.

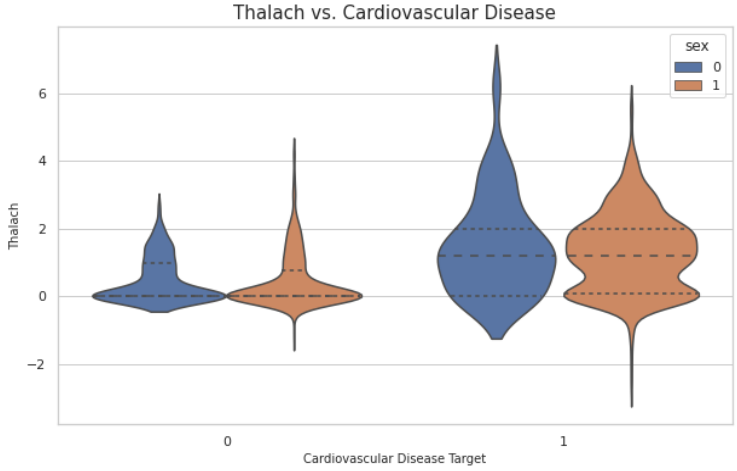

Fig. 7 Cardiovascular Disease Target with Thalach

The box plot was used to see the differences between the levels of ST depression, In Fig 8. class=0 (positive) people experienced an elevated range, while class=1 (negative) people experienced a lower range. No major disparity was observed between the genders.

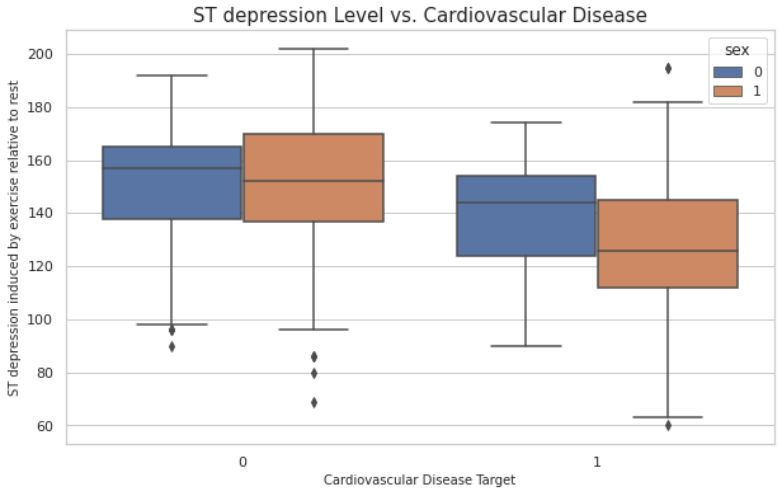

Fig. 8 Cardiovascular Disease Target with Exang

When comparing both the classes, there is a huge contrast in the means for many of the features mentioned in Fig 2. Examining both box and violin plots, the inference was the positive patients experience an elevated heart rate i.e., Thalach, and also exhibit about 33.3 % of exercise relative to rest i.e., Oldpeak.

3) Classification

The Features specified in Fig 1. provide information to diverse ML algorithms such as Gradient boost, XG boost, Random Forest, and K-nearest classification techniques. Firstly, the dataset splits into two, training and the testing. Usually, to train a model training dataset is applied, and a testing dataset is utilised to control the execution of the trained model. Each algorithm undergoes various parameters and performance is calculated. The algorithms that are investigated in this paper are classified below.

Gradient Boosting

Gradient boosting algorithm has marked applications across various technological domains but here it is focused on medical purposes. In the gradient boosting algorithm, prediction accuracy is increased through improving multiple models in classification by putting importance on these training events that are challenging to evaluate. This model minimizes the loss function constantly by using the Gradient Descent method so that the inferring the difference from the true value of the patient's data and Eq.1 have estimated the result values for each leaf (q) in the tree.

$$A = \frac{Total\ number\ of\ residuals}{sum\ of\ each\ n(1-q) for\ each\ sample\ in\ the\ leaf} \qquad (1)$$

Random Forest Classifier

In this analysis, the classifier arranges a tree for the data and tries to predict the output on that. Disease prediction was made by considering the mean of the output from various trees. Here in this classifier the n_estimator = 5 was fixed along with the random_state = 1. Criterion given in Eq.2 is a function to study the accuracy of the split, and here it has been set to 'entropy'.

$$\text{Entropy} = \sum_{c=1}^{a} - n(qi)\ log2\ (p(qi)) \qquad (2)$$

Where n(*qi*) is the probability of the class *qi* in a node.

K-Nearest Neighbor

The principal reason for using the K-Nearest Neighbor algorithm in predicting cardiovascular disease is because it works by assuming the correlation between the unique and possible patient data and introducing the new data into the category that is most similar to the possible class/group. K-NN algorithm can be used for both regression and classification problems. Whenever new patient data come into the dataset then it can be easily classified into a suitable group by finding the perfect k-value. To determine the perfect value for "K" is by trying out some random values until finding a suitable one. The parameters that were used in this classifier were n_neighbors=5 and the metric='minkowski' (Eq.3) was to determine the distance between the points. The minkowski metric is mainly used to find distance similarity and shown below.

$$Minkowski = (\sum_{o=1}^{e}(|ai - bi|)^{a})^{1/a} \quad (3)$$

Where a denotes the order of the model, if order a= 1 it denote Manhattan, and when order a=2, will denote Euclidean.

XG Boosting

It is a structure designed for performance and speed. In this research, predicting the problems that involve unstructured patient data (images, text, etc.) Gradient boosting works where the new models are created that calculate the error in the preceding model. Finally, the residual data is accounted for making the final predictions with the accumulated patient's data. The main reasons for the great predicting performance of the algorithm are as follows, cross-validation, missing value, save and load, regularization, flexibility.

4) Important features for the risk of CVD

Among the 12 features compared, Fig 10. shows the top few features that are contributing to the risk of cardiovascular disease. It can be done by taking the highest performing model and enumerating the features into it. The results depict the level of features and can be concluded with the last 4 features because their values are higher.

Exercise induced angina	Exang
Peak Exercise ST segment	Slope
Chest Pain	Cp
Exercise induced at rest	Oldpeak

Fig. 10 Important features

Result and Analysis

The final results from the boosting and bagging algorithms are acquired and analyzed further by using a classification report and confusion matrix. Further, the stacking method was used to exhibit an enhanced accuracy followed by the algorithms. In the classification report, the metrics used to follow the predictive analysis are as follows, Accuracy, Recall, Precision, F-1 score. The metrices and its calculation are shown in below graphs. A **true positive** is a decision where the model *accurately* prophesies the *positive* inmates. A **true negative** is a decision where the model *perfectly* foretells the *negative* inmates. A **false positive** is an event where the model *incorrectly* divines the *positive* inmates. A **false negative** is a decision where the design *unfairly* guesses the *negative* inmates.

Precision

The below graph is the calculated precision of the machine learning models used in this research. It's the proportion of perfectly predicted positive values to the entire predicted positive values. Eq.4 describes the formulation of precision. The inference observed in this graph is the highest percentage of negatively predicted patients is in XGBoost. Surprisingly, the positively predicted percentage were in the K-NN algorithm with 95% in Fig 11.

$$TP / TP + FP \qquad (4)$$

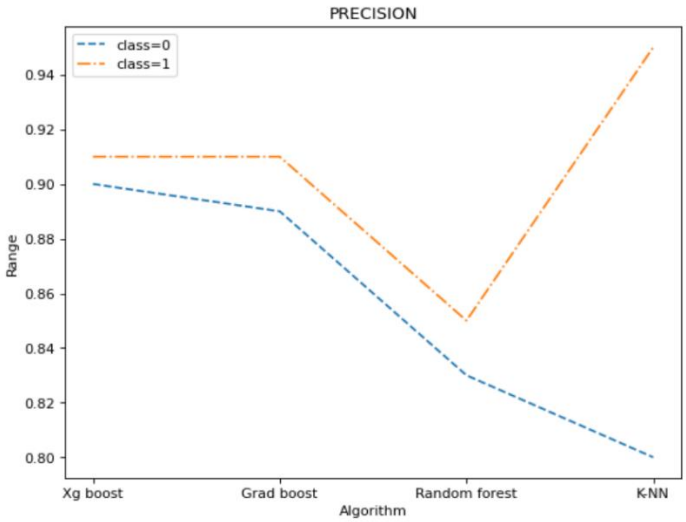

Fig. 11 Precision values of algorithms'

Recall

It is a relationship of correctly evaluated positive values to every observation in the original value. Eq.5 describes the formulation of recall. The calculated recall values for machine learning models are integrated into the graph as shown. Taking a closer look at the recall's graph shows the negatively predicted patients in K-NN holds leading among other algorithms with a percentage of 95% but in the positively predicted part XG6 boost holds the highest of 90% in Fig. 12.

$$TP / TP + FN \qquad (5)$$

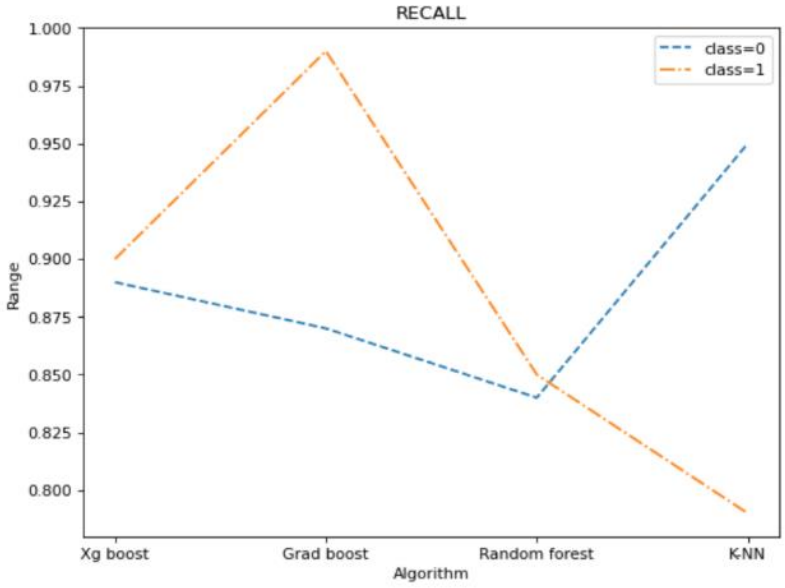

Fig. 12 Recall values of algorithms'

F-1 Score

It is an angled mean of precision and recall. Consequently, this score exerts both false negatives and false positives within the record. Eq.6 describes the formulation of F-1 score. The calculated f-1 scores for machine learning models are integrated into the graph as shown. The inference observed in the F-1 score's graph (Fig 13.) while predicting the negative patient both the boosting algorithm stands the highest of 90% at the same time, in the positively predicted segment XGBoost holds the crown with 91%.

$$(TP + TN) / TP + FP + TN + FN \qquad (6)$$

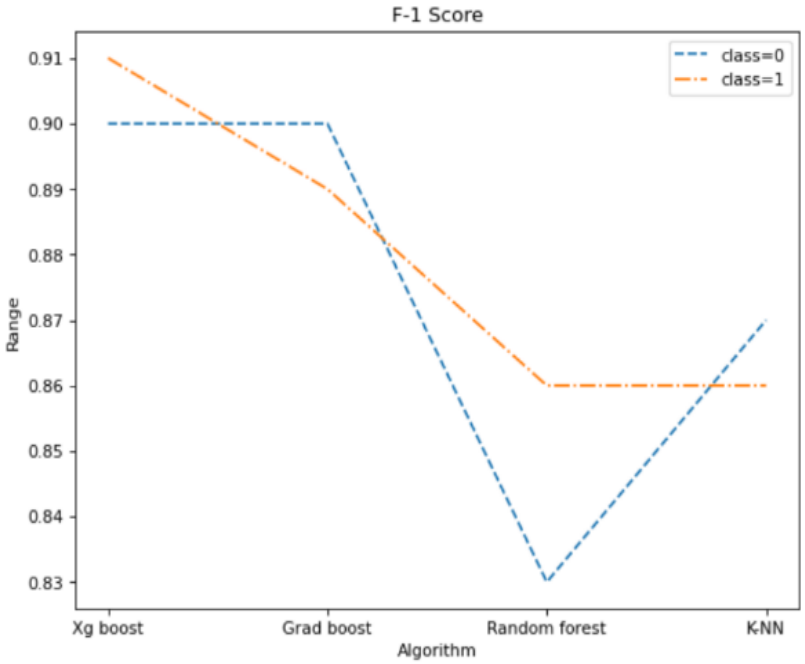

Fig. 13 F-1 score values of algorithms

In Table 1, Class is defined as the diagnosis of cardiovascular disease (predictor). For instance, 0 as patient with negative risk and 1 as patient with positive risk.

Accuracy

It is a usual spontaneous review model, and totally a proportion of accurately forecasted values to the entire values given in Fig 14. As the metrics graphs show that the most positively predicted percentage were in the XGBoost. Thus, XGBoost has the highest overall percentage of 90 followed by Gradient boosting with 89%, Random Forest with 84% and K-NN with 87%.

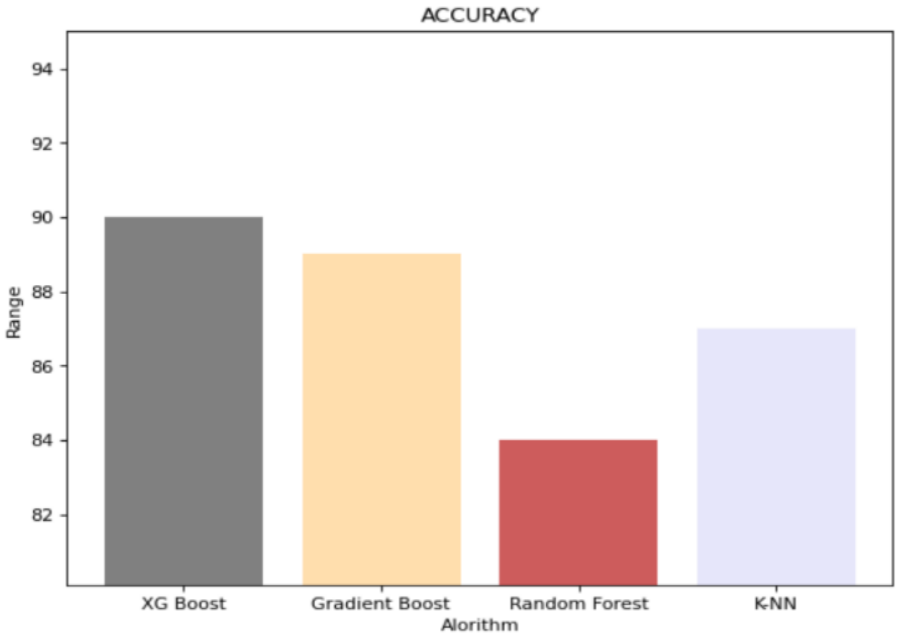

Fig. 14 Performance of the classifiers before stacking

The classification report for each machine learning model is calculated and found out the accuracy. XG boost algorithms give 90% accuracy and which was the highest among all others that compared given in Table 2. Followed by another boosting algorithm that gives 89% accuracy i.e., Gradient boosting algorithm. Increasing the number of trees by allowing bagging algorithms like Random Forest to give out more accuracy.

Stacking and Enhanced Result

The stacking of algorithms is a method of combining individual algorithms to provide an increased efficiency i.e. Stacking involves **practicing various base patterns to determine the objective in a machine learning model.** The given meta classifier was RFC. In this paper, processing of all the base learners produces an improved accuracy of 92% (Table 1.) by the use of stacking algorithm, previously the accuracy of the XGBoost algorithm was the highest among all. Fig 15. Provides a compared accuracy chart.

Table 1 Predictive analysis of algorithms

ALGORITHM	Accuracy	Recall		Precision		F1-Score	
		0	1	0	1	0	1
Gradient Boost	**89%**	**0.87**	**0.89**	**0.89**	**0.91**	**0.90**	**0.89**
Random Forest	**84%**	**0.84**	**0.85**	**0.83**	**0.85**	**0.83**	**0.86**
K-Nearest Neighbor	**87%**	**0.95**	**0.79**	**0.80**	**0.95**	**0.87**	**0.86**
XG Boost	**90%**	**0.89**	**0.90**	**0.90**	**0.91**	**0.90**	**0.91**
Stacking	**92%**	**0.93**	**0.92**	**0.91**	**0.94**	**0.91**	**0.93**

The Predicted values and actual values of XGBoost, Gradient boost, Random Forest and K-nearest neighbor can be found in Table 2.

Table 2 Predicted and actual values of ML model

ALGORITM	TRUE POSITIVE	TRUE NEAGATIVE	FALSE POSITIVE	FALSE NEGATIVE
Gradient Boost	**97**	**116**	**14**	**11**
Random Forest	**93**	**108**	**18**	**19**
K-Nearest Neighbor	**106**	**100**	**5**	**27**
XG Boost	**99**	**116**	**12**	**11**
Stacking	**145**	**130**	**10**	**13**

The Fig 15. represents the compared accuracy of all the algorithms and the stacking of XGBoost, Gradient boost, Random Forest, and K-nearest neighbor with a meta-classifier RFC, shows an increased accuracy from 90% (XGBoost) to 92% (Stacking).

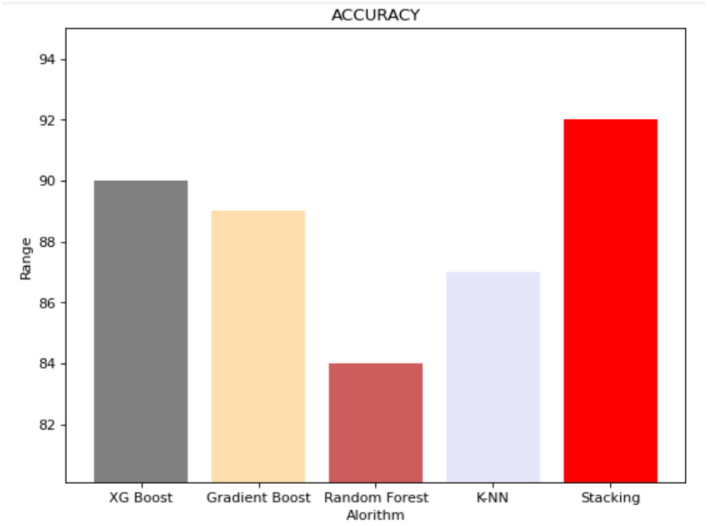

Fig. 15 Performance of the classifiers after stacking

Conclusion

To recapitulate, raising in mortality due to cardiovascular diseases has become essential to produce a method to prognosticate cardiovascular conditions efficiently and precisely. Cardiovascular disease emerging as the sole cause of mortality in many countries. The immense hardship caused, in terms of suffering and expenses on health care, is intensifying. So, the desire toward the research is to gain a productive machine learning algorithm with the tuned parameters and enhanced results for the early discovery in risk of heart disease. The research first analyses the user-level security to identify the user who wants to access the data and decrypt it accordingly, then compares the accuracy of each of the machine

learning models. The results exhibit that the XG boost algorithm is effective with 90% accuracy followed by the Gradient boost algorithm 89% and to enhance the overall performance of the model, the stacking method was used, it produces an improved accuracy of 92%. The machine learning models used in this paper are limited to a certain extent by allowing a maximum depth of the tree. Meanwhile, without limiting the tree nodes, the bagging algorithms like Random Forest and K-nearest neighbor would have higher accuracy than the boosting algorithm. Future works on this paper can be enhanced by even restricting the parameters to find risk of cardiovascular disease. Though this acts as a principal objective for the medical industry, developing it further screening for every patient who is diagnosed with a heart-related disease so that this research serves a noble purpose.

References

Sulabha S, Chaitrali S. Dangare, (2012) "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" *International Journal of Computer Applications (0975 – 888)* Volume 47– No.10, June 2012.

Niti Guru, Anil Dahiya, Navin Rajpal, (2007) "Decision Support System for Heart Disease Diagnosis Using Neural Network", *Delhi Business Review,* Vol. 8, No. 1. [Jan 2007 - Jun 2007].

Dundigalla Ravi, Avi Agarwal, Poonam Ghuli, (2020) "Heart Disease Prediction using Machine Learning" *International Journal of Engineering Research & Technology (IJERT),* ISSN: 2278-0181, Vol. 9 Issue 04, April-2020.

Fahd Saleh Alotaibi, (2019) "Implementation of Machine Learning Model to predict Heart Failure Disease", (IJACSA) *International Journal of Advanced Computer Science and Applications,* Vol. 10, No. 6, 2019.

Ramya G Franklin, Anjan Nikhil Repaka, Sai Deepak Ravikanti, (2019) "Design and Implementing Heart Disease Prediction Using Naive Bayesian" *International Conference on Trends in Electronics and Informatics (ICOEI),* 23-25 April 2019.

J. Thomas, and R. Theresa Princy, (2016) "Human heart disease prediction system using data mining techniques." *International conference on circuit, power and computing technologies (ICCPCT). IEEE,* 2016.

Lee B.J., Lee H.G., Noh K., Shon HS., Ryu K.H., (2006) "Cardiovascular Disease Diagnosis Method by Emerging Patterns", *International Conference on Advanced Data Mining and Applications, ADMA 2006: Advanced Data Mining and Applications, 4093,* 819-826.

Yang Q, Zhang Z, Khoury MJ, Moonesinghe R., (2019) "Prevalence and cardiovascular health impact of family history of premature heart disease in the United States". *Journal of the American Heart Association*. 2019;8(14): e012364.

Solanki, Y.S. Chakrabarti, P. Jasinski, M. Leonowicz, Z. Bolshev, V. Vinogradov, A. Jasinska, E. Gono, R. Nami, (2021) "Hybrid Supervised Machine Learning Classifier System for Breast Cancer Prognosis Using Feature Selection and Data Imbalance Handling Approaches". *Electronics 2021,* 10, 699.

M. Omair Shafiq, and T. Obasi, (2019) "Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack and Diseases," *2019 IEEE International Conference on Big Data (Big Data)*, 2019, 2393-2402. http://doi.org/10.1109/BigData47090.2019.9005488

T. Mahboob, R. Irfan and B. Ghaffar, (2017), "Evaluating ensemble prediction of coronary heart disease using receiver operating characteristics," *2017 Internet Technologies and Applications (ITA)*, 2017, 110-115, http://doi.org/10.1109/ITECHA.2017.8101920

Prabhat Pandey, Atul Kumar Pandey, & K.L., (2014)" Classification Model for the Heart Disease Diagnosis" Jaiswal Global Journal of Medical research: F Diseases Volume 14 Issue 1 Version 1.0 the Year 2014, *Double-Blind Peer Reviewed International Research Journal, Global Journals Inc. (USA)* Online ISSN: 2249-4618 & Print ISSN: 0975-5888.

Bashir, Saba & Qamar, Usman & Javed, Muhammad, (2014). An Ensemble based Decision Support Framework for Intelligent Heart Disease Diagnosis. *International Conference on Information Society, i-Society 2014.* http://doi.org/10.1109/i-Society.2014.7009056

R. Delshi Howsalya Devi, G. Karthiga, C. Preethi, (2014) "Heart Disease Analysis System Using Data Mining Techniques", *2014 IEEE International Conference on Innovations in Engineering and Technology (ICIET'14),* Volume 3, Special Issue 3, March 2014.

Sellapan palaniappan, Mat Ghani, Ts. Dr. Mohd & Awang, Raflah, (2008). "Intelligent heart disease prediction system using data mining techniques". 8. 108 - 115. http://doi.org/10.1109/AICCSA.2008.4493524.