COMPUTOLOGY: Journal of Applied Computer Science and Intelligent Technologies
Volume 3, Issue 2, Jul-Dec 2023, pp. 133-152
DOI: 10.17492/computology.v3i2.2307
https://www.journalpressindia.com/cjacsit

Forecasting the Risk of Coronary Heart Diseases using Machine Learning Algorithms

Lakshmi J.V.N.[1] *and Anirban Das*[2]

ABSTRACT

There are many emerging technologies such as machine learning and data analytics that offer promising solutions to healthcare challenges, biomedical communities, and patient care. Early detection of disease symptoms helps improve disease management strategies. Early detection also helps with disease symptom control and efficient therapy. In this study, we present a complete preprocessing strategy to predict coronary heart disease (CHD). The method includes computing null values, standardizing, categorizing, normalizing, resampling, and finally predicting. The purpose of the study is to predict CHD using machine learning techniques such as Random Forest, k-nearest neighbors, decision trees, logistic regression, and gradient boosting. We propose K-fold cross validation to provide predictability across the data. We test these algorithms on 4240 records from the "Framingham Heart Study" dataset. We also use a feature selection algorithm to reduce the dimensionality problem while keeping the computational complexity close to acceptable accuracy. The feature selection algorithm reduces the dimensionality problem and keeps the computational complexity close to acceptable accuracy. To predict accurately the risk of heart disease and to assess if a person has a risk of CHD, a new ensemble method using gradient boosting, random forest and k-nearest neighbor with majority vote has been tested with 96.16% accuracy and 0.96 ROC-AUC score. The experiments show that advances in machine learning, combined with predictive analytics, offer a potential environment for finding intelligent solutions, showing the potential of prediction in the field of cardiovascular disease and beyond.

Keywords: *Coronary Heart Disease; Gradient Boosting; Random Forest; KNN; Early detection; Machine Learning.*

1.0 Introduction

According to the World Health Organization (WHO), heart disease and Stroke (CVD)

[1]*Corresponding author; Department of Computer Science, Patel Institute of Management and Sciences, Bangalore, Karnataka, India (E-mail: jupudilakshmi@gmail.com)*
[2]*LJMU, Bangalore, Karnataka, India*

caused 17.9 million deaths worldwide in 2019, representing 32% of global fatalities, and 80% of these deaths were caused by artery blockage and brain stroke. If precautions are not taken properly, this ratio is expected to increase (American Heart Association, 'heart disease and stroke statistics-2021', 2021) [1]. Risk factors for heart disease include high cholesterol, heavy smoking and dietary habits, excessive coffee and alcohol use, high stress levels, high blood pressure, obesity, and family history of illness [2].

Cardiovascular diseases are very common in low- and middle-income countries [3]. Heart disease causes pain in arms and chest. Examples of cardiovascular diseases include coronary heart disease (CVD), cerebrovascular disease (CVI), peripheral arterial disease (PAD) and congenital heart disease (CJD) [4]. Although advances in health care and medical research have shown a linear increase in the rate of CHD over time, researchers around the world have been trying to identify the factors associated with future CHD risk [5]. Resent methods for the prediction and diagnosis of cardiac disease are based on the patient's medical history, age and gender, heart rate and physical reports. In some cases, medical specialists can accurately predict a patient's cardiac problems up to 67% [6] [7]. All these procedures add up to postponing diagnosis tests which often leads to incorrect diagnosis and mechanical failure which lowers the accuracy. It is expensive and involves complex calculations and evaluations which take long time to complete [8] [9].

The medical industry needs an automated, intelligent, and robust model that predicts heart disease and assists in decision-making [10]. This model can be achieved by combining large datasets of patients with machine learning (ML) or deep learning (DL) algorithms and smart decision-making systems [11]. If properly analysed, the massive data available in healthcare database repositories can help reduce the prevalence of these disorders [12]. This will make it simpler and faster for physicians to predict ailments and diagnose them. Such massive data needs to be evaluated to develop an effective disease management plan.

Artificial intelligence based early disease detection, severity grading and prediction are the common occurrences [13]. This will help delay disease onset, enhance patient quality of life (QOL), and reduce medical costs. Continuing in this vein, machine learning models are used to create classification models by combining the appropriate features of each dataset [14]. The motivation of this research is to evaluate and quantify the usefulness of a few machine learning models for the prediction of coronary heart disease (CHD) [15] [16].

The issue of class imbalance affects the accuracy of the classification algorithm. Furthermore, when the data has an imbalance, many features need to be used for

prediction [17]. This significantly increases the complexity of the solution, making it unsuitable to be used in a real-world setting. Existing feature selection algorithms need to be improved to decrease the computational complexity while maintaining the accuracy. In this study, we will attempt to develop classifiers for coronary heart disease prediction using random forest, k-nearest neighbor, decision tree, logistic regression, gradient boosting, and K-fold cross validation. We will also test a new ensemble method (majority voting) to see if these models can accurately predict the chances of coronary heart disease and identify the risk factor for a person concerned with coronary heart disease.

The research proposes an ensemble method with majority voting to aid in the correct prediction and analysis of patients with heart disease. This decision support system will be more suitable for the analyzing of heart disease and thanks to machine learning algorithms. The method is non-invasive, agile, rapid, reproducible and target-oriented and is applied to screen large numbers of patients internationally based on clinical knowledge which can easily be acquired in medical clinics. The study's most important feature is that it can connect and address everyone who is affected by this problem. Additionally, the computational time is predicted to decrease which will be beneficial during deployment.

2.0 Literature Review

2.1 Data analytics in healthcare

The authors of [18] [19] proposed that big data and analytics are two perspectives within the term BDA. Volume (due to the large amount of data), velocity (due to real-time and fast accumulation of data), and variety (due to the various forms of data) are the three features of big data in the healthcare industry (structured data, unstructured data, semi-structured data). The newest addition to big data, veracity, is the improvement of data dependability as data is electronically and computer generated from handwritten notes. The term "analytics" encompasses management information systems [MIS], statistics and operational research [OR]. It's a combination of business intelligence reporting, advanced statistical techniques in data mining, forecasting and descriptive analysis, as well as other OR techniques such as simulation and optimization. ODR has benefited from advanced problem solving and improved decision-making from the processing and analytics of big data.

2.2 Data mining in diagnosis of chronic diseases

In their research [20], the authors looked at the limitations of traditional health administration and the state of disease risk prediction model research in my nation. In a

big data health management application, a disease early warning model is built to evaluate the risk factor and suffering of a patient from a particular condition to identify the disease and its risk. Factors that influence when early intervention is needed to prevent disease or control its progression. Because of the rapid development of medical information, which contains a lot of future knowledge, huge amounts of medical data have been generated.

From clinical research, medical decision-making, and early warning of diseases this application could revolutionize chronic diseases diagnosis and treatment. Medical data is a database of medical and healthcare information, including patient records, imaging, gene test, epidemiological survey, and so on. A large part of medical data is stored electronically as medical technology [21].

One sort of medical data is clinical data A database containing medical and healthcare information such as patient records, imaging data, gene test data, epidemiological survey data, and so on is known as medical data. A significant amount of medical data is being kept in electronic form as medical technology [22].

2.3 Data mining in prediction of cardiovascular heart disease

As we live in a post-modern era, our daily lives are undergoing significant changes that have both positive and negative impacts on our health. As a result of these changes, the incidence of many diseases has increased dramatically. Lives are in danger. Blood pressure changes, pulse rate changes, sugar levels, and many other factors can lead to cardiovascular disorders such as narrowing or clogging of the arteries which can lead to heart failure, heart aneurysm, peripheral artery diseases, heart attack, stroke, sudden cardiac arrest, and many other types of heart diseases. Different medical tests that consider family history and other factors are used to diagnose various types of heart disease. However, it is very difficult to predict cardiac problems without carrying out any medical tests.

As mentioned in [23], to diagnose heart disorders and to take all necessary steps to prevent it at the earliest possible time at an affordable cost, ‘data mining’ is used for the prediction of heart disorders. The pre-measurements and studies obtained from this technology determine the probability of discovering heart illnesses which can be completely cured with the proper diagnosis

2.4 Predictive analytics in heart diseases

The authors [24] looked at different research techniques used to predict heart disease. The severity levels of the disease were justified using different methods, such as K-nearest neighbour algorithm (KNN), genetic algorithm (GA), decision trees (DT) and

naive bayes (NB), support vector machines (SVM), particle swarm optimization (PSO), artificial neural networks (ANN), and random forest (RF). It is important to use medical data and information to its fullest extent, as well as to store and retrieve it. Clinical data health informatics techniques are used to consume and store data. This method will help to gain a better understanding of problem solving and making correct decisions. There has been a lot of progress in health care technology for collecting, treating, communicating, and researching. A comparative study of efforts to predict heart disease was conducted between 1990 and 2019 [25] [26]. Machine learning was used by almost 40% of researchers, and artificial neural networks were used by the majority [27].

3.0 Research Methodology

A high-precision machine learning model that accurately detects heart related diseases on patient medical data. Data pre-processing includes outlier removal, imbalance class handling and replacement of missing values. Classification using an ensemble of random forest, gradient boosting and KNN. Selection of features using feature-sensitivity technique. The proposed method uses fewer features and reduced computational complexity. The data is mostly searched for likely outliers in the pre-processing phase. We will use the Framingham dataset as it contains all of the factors needed for disease prediction.

3.1 Data selection

The used dataset that's a subsection of the Framingham Heart Study (FHS) dataset, and it's accessible for free through the Framingham Heart Institute.

There are 4240 records in the available segment of the FHS dataset. The data comes from a long-term study of a Framingham, Massachusetts population. The research focuses on the causes and origins of cardiovascular disease, and it falls within public health management domains [27]. The Framingham Heart Study aimed to discover the risks that influence a person's health with coronary heart disease. There are 16 different features in the dataset that affect coronary heart disease is given in Table 1.

Table 1: Attributes of the Dataset and its Interpretation

Attribute	**Interpretation**
gender	Female : 0; Male : 1
Age	Age at the examination time
education	1: high school

	2: high school or GED 3: college or vocational school 4: college
currentSmoker	0 = nonsmoker; 1 = smoker
Diabetes	0 = No; 1 = Yes
totChol	Total cholesterol inside patient's body (mg/dL)
sysBP	Systolic Blood Pressure (mmHg)
diasBP	Diastolic Blood Pressure (mmHg)
cigsPerDay	Number of cigarette smoked per day (average)
BPMeds	Is the person on BP medicines
prevalentStroke	If the person had any prevalent stroke
prevalentHyp	Any beneath prevalent
BMI	Body Mass Index: Weight (kg) /Height(meter-squared)
Heartrate	Beats/Min (Ventricular)
glucose	Amount of glucode in mg/dL
TenYearCHD	Risk of developing CHD (Yes:1; No:0)

3.2 Data preprocessing

Basic descriptive statistics tables, skewness, and other descriptions such as min, max, percentile values, and mean were computed rst step in data preprocessing.

Table 2: Count of Missing Values

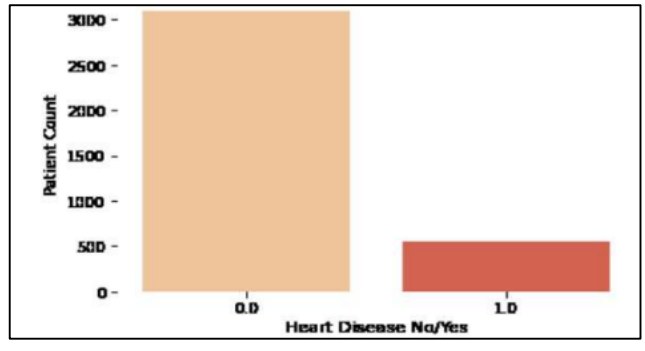
Attributes	Count of Missing Values
Male	0
Age	0
Education	105
CurrentSmoker	0
CigsPerDay	29
BPMeds	53
PrevalentStroke	0
PrevalentHyp	0
Diabetes	0
totChol	50
SysBP	0
diaBP	0
BMI	19
Heartrate	1
Glucose	388
TenYearCHD	0

It also contains missing value detection and removal, as well as the conversion of categorical data (the sex, Current smoker, and diabetes columns) to integers. The mean values of each column were substituted for the missing values in cigsPerDay, totChol, BMI, glucose, and heartrate. Also deleted from the dataset were the missing values of BPMeds, which are categorical, and education (ordinal with range 1-4). The number of missing values identified in each feature or attribute is shown in Table 2 [27].

3.3 Class balancing

Researchers have encountered imbalance of class as a serious issue disturbing accuracy in numerous machine learning problems [28]. When the samples from different classes are unequal, a problem arises. The class imbalance problem also exists in the Framingham dataset. In the Framingham dataset, there are 644 examples of class 1 and 3596 samples of class 2, indicating the dataset is sufficiently imbalanced, as seen in Figure 1 Using the 'Synthetic Minority Over-sampling Technique,' the suggested approach addresses this specific issue (SMOTE). Also, the suggested framework uses SMOTE to make the samples in both the classes of the Framingham dataset equal [27].

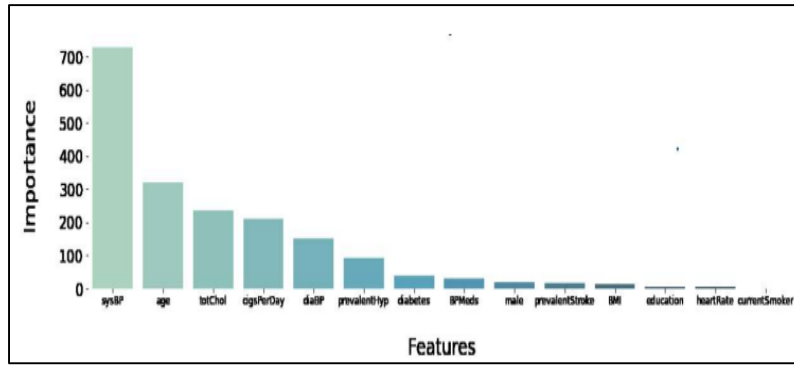
Figure 1: Class Imbalance



3.4 Feature selection

Selection of Features has become a critical component of machine learning, particularly for datasets with many features and samples. The relevant characteristics are chosen in feature selection so that the algorithm's efficiency improves as the computational time and complexity reduce. As a result, selection of features is the crucial phase in machine learning, and the algorithm's correctness is strongly reliant on it. Because the Framingham dataset has 15 qualities, the approach proposes selecting the most important ones. The technique of 'feature importance' is applied in the proposed framework. The outcomes of this technique are depicted in Figure 2 [27].

Figure 2: Important Features



By removing the duplicate feature from the data, feature importance lowers overfitting. Because the data chosen is not duplicated or misleading, it contributes to increased accuracy. In addition, because less data is necessary for training, the suggested technique is computationally inexpensive. The feature importance score assigns a value to each data feature; the higher the score, the more essential or useful the feature is for estimation. Feature significance is an intrinsic class in Tree based classifiers; the suggested framework extracts most significant features using the Select K Best class. Increases in impurity of node are biased with the chance of reaching that node to evaluate the relevance of a trait. Equation (1) is used to calculate the probability of a node [27].

$$Node\ Probability = \frac{Number\ of\ samples\ reaching\ that\ node}{Total\ Number\ of\ samples} \quad \ldots(1)$$

The node probability value determines the feature's importance. Once the value is high, the more significant is the feature. The features having a score of more than 100 are chosen for more analysis.

Table 3: Selected Features

Attributes	Score
SysBP	727.935535
Age	319.266019
totChol	235.502392
CigsPerDay	209.897040
diaBP	152.748563
SysBP	727.935535

Table 3 lists the significant features extracted from the dataset. The attribute 34 'Education' has been carefully deleted from the dataset because it has little impact on CVD prediction. Table 3 slected Features Two more algorithm is also proposed for better results 1) Relief Feature Selection Technique and 2) Least Absolute Shrinkage and Selection Operator Algorithm (LASSO) [27].

3.5 Methods applied

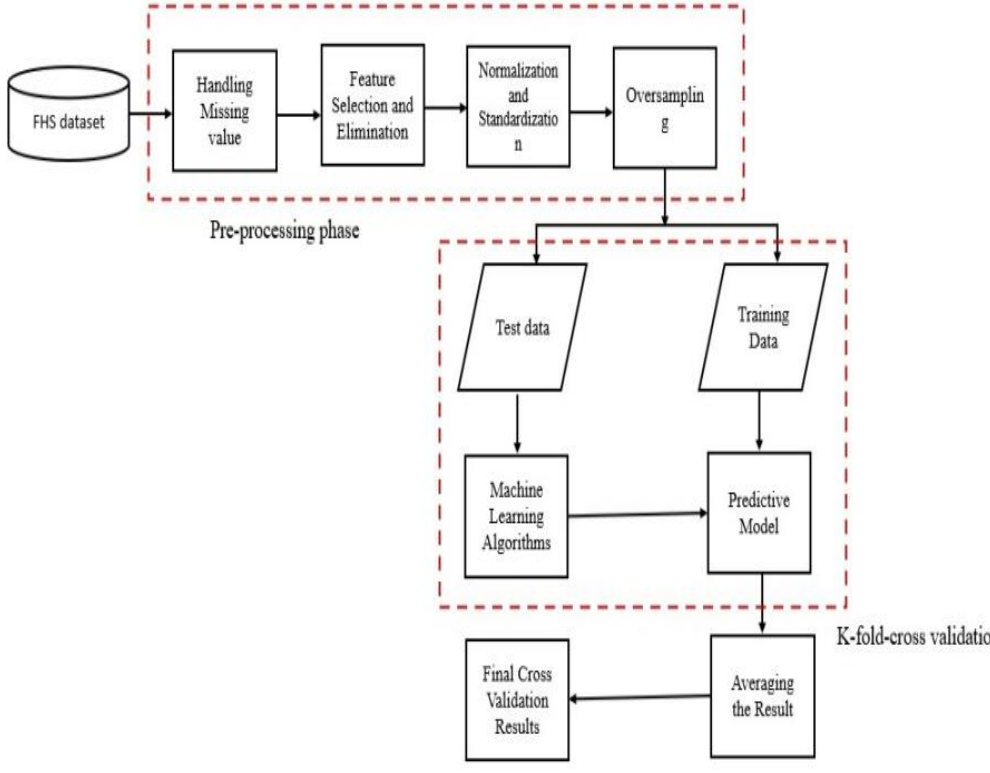
Ensemble learning is a useful approach for refining overall model performance by combining the forecasts of many learning algorithms [29] [30]. Ensemble learning is the best option [31]. Some models may operate well with data, while others may not. When we combine these models, their flaws are eliminated. For the ensemble approach, the top 3 classification methods are chosen. For accurate heart disease prediction, the proposed model ensemble k-nearest neighbour technique, Random Forest and Gradient Boosting to be used. To improve wise the predictions, hyperparameter tuning is also proposed then a majority vote or hard vote to be used. Comparative study of the evaluated model performances from previous research shown below in Table 4. The proposed method is applied on each test case and the ultimate result is determined by a majority vote which is validated by k-fold cross validation as shown in Figure 3 [27].

Table 4: Comparative Study of the Methods

Methods Used	Accuracy achieved	References
Decision Tree (DT), K-Nearest Neighbor (KNN), and Random Forest (RF)	Validation accuracy of DT, KNN and RF are 97.17%, 99.34% and 90.92% respectively	P. Ghosh et al
K-nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), Naïve Bayes (NB) and Random Forest (RF)	72.94%, 88.65%, 78.41%, 72.12%, 65.54% and 94.63% respectively	Kwakye Kelvin and Dadzie Emmanuel
Supervised models such as AdaBoost (AB), Decision Tree (DT), Gradient Boosting (GB), K-Nearest Neighbors (KNN) and Random Forest (RF) together with hybrid classifiers are applied	DT-86.97%, RF-88.65%, KNN-83.61%, AB-89.07% and GB-86.97%	Pronab Ghosh et al
Artificial Neural Network (ANN), AdaBoost and Decision Tree (DT)	ANN-98%; AdaBoost-90.18% and DT-93.18%	Terrada Oumaima et al
DT, Naïve Bayes (NB) and Support Vector Machine (SVM)	DT-70.9%; NB-71.8% and SVM-71.02%	Amanda H. Gonsalves et al
Random Forest (RF), Decision Tree (DT) and K-Nearest Neighbors (KNN)	RF-96.80%, DT-92.45% and KNN-92.81%	Krishnani D et al
Ensemble of RF, GB and KNN		Proposed

A critical stage in the research process is to choose a relevant dataset. After choosing the data, it must be appropriately collected and structured. The Framingham Heart Institute provided a free download and collection of the Framingham Heart Study (FHS) dataset. To clean the data, the data was first pre-processed. Missing values and class imbalance are checked via additional data processing. The Outliers detected were excluded to reduce the noise in the data and to maintain the appropriate model's accuracy. In feature selection, the relevant properties are chosen so that the algorithm's efficiency improves as the computing time and complexity decrease.

Figure 3:Proposed Ensembled Methodology



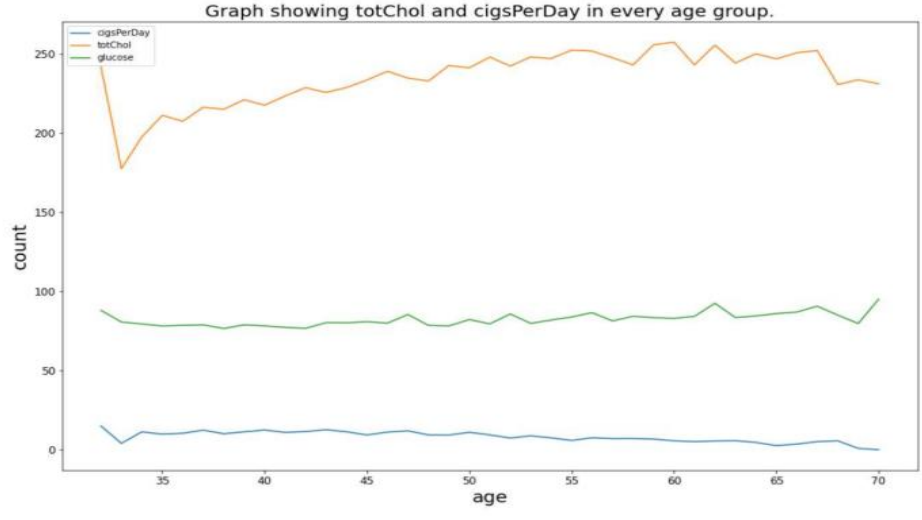
The model is then built using clean data and features that have been chosen. Before deciding on a final model, measures such as accuracy, loss, recall, and precision will be examined. The chosen model will be tested as an ensemble and compared to previous models to determine how well the new model performs. A good model can

benefit society by assisting people in maintaining their fitness and having a healthy heart.

4.0 Exploratory Data Analysis

In Machine Learning, analysis is a process of delving into the data and learning distinct properties, frequently using visual means. It enables a better understanding of data and the discovery of interesting patterns within it. Before performing analysis and running the data through an algorithm and through understanding of the data is critical. Data patterns are frequently necessary to determine the essential variables that define the appropriate outcome. Recognizing data flaws is also necessary for getting the accurate prediction. After doing an initial data check, a full knowledge of the data is required to the model to produce more reliable and accurate results. Any recurrent patterns and strong correlations thar exists need to be understood. Exploratory Data Analysis is the process of learning everything there is to know about a set of data. It's an important component to deal with the data, and it's only possible with exploratory data analysis (EDA). It aids in the gathering of insights and the better understanding of data, detecting anomalies and unneeded numbers. A machine learning model are incorporated to predict more accurately, resulting in more precise outputs. And finally aids in the selection of a superior machine learning model.

Figure 4: Relationship between totChol, cigsPerDay and age

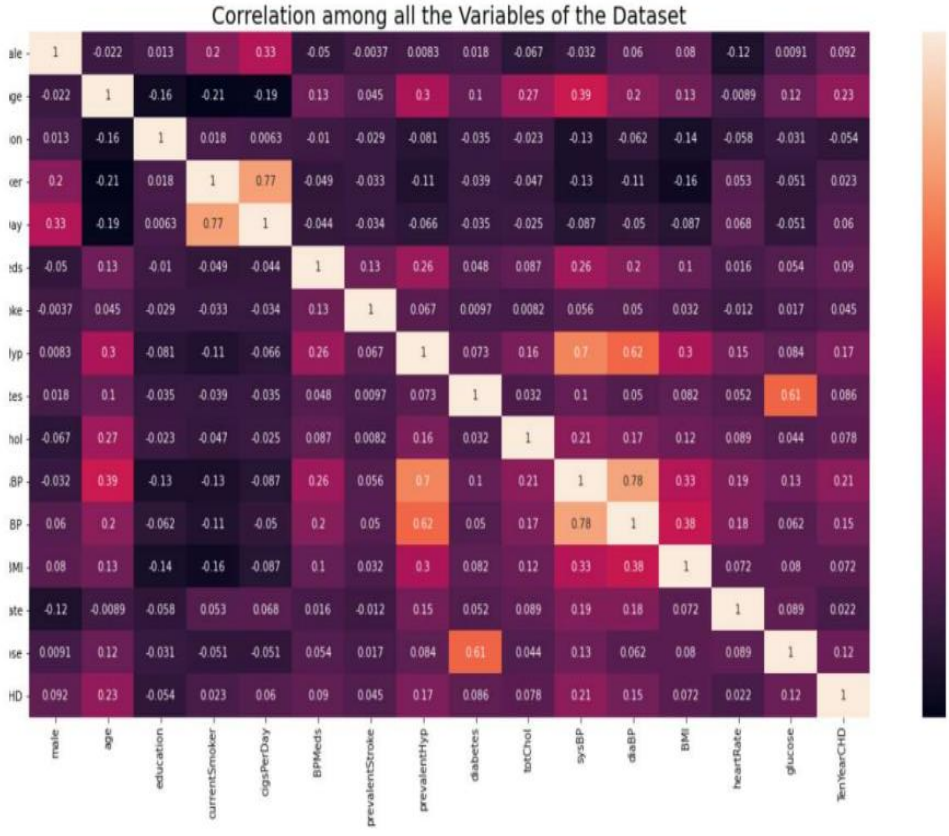


Checking on the relationship between age and cigsPerDay, totChol, glucose it is observed TotChol and glucose have a minimal relationship. For lower age ranges, totChol has a steep, linear, and inverse graph, while cigsPerDay has a reasonably parallel relationship with age. Same is shown in Figure 4 describing the multivariate analysis.

The correlation among all the variables in the dataset. From the observations the correlation coefficient between education and the goal variable TenYearCHD is very low and even negative when compared to all the independent data as shown in Figure 5.

The dataset study revealed demographical, medical, and behavioural factors predict a patient's CHD risk. Missing values of the dataset are treated using mean. Exploratory Data Analysis determines the relationship between the predictor factors and the target variable. The significant data imbalance that was addressed by resampling of positive cases.

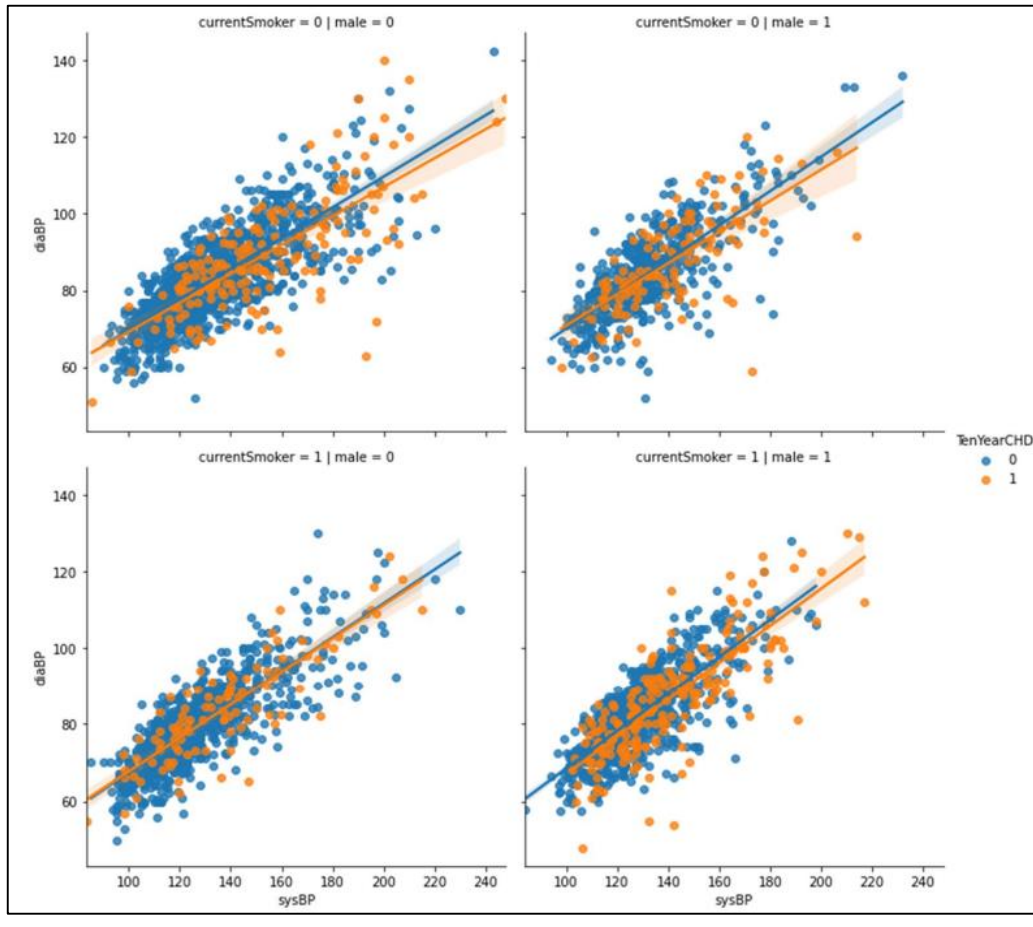
Figure 5:Correlation among all the variables



Feature selection is also used to improve model performance and reduce computing costs. Scaling is applied to change the data before it was separated into the training set and test set. This Analysis aided in the critical process of conducting preliminary investigations identifies patterns, anomalies, test hypotheses, and validate assumptions using summary statistics and graphical representations.

Figure 6 shows the association between systolic and diastolic blood pressure for patients based on their gender and current smokers and plots the best fit line.

Figure 6: Distribution of sysBP vs diaBP with respect to current Smoker and gender



5.0 Results and discussion

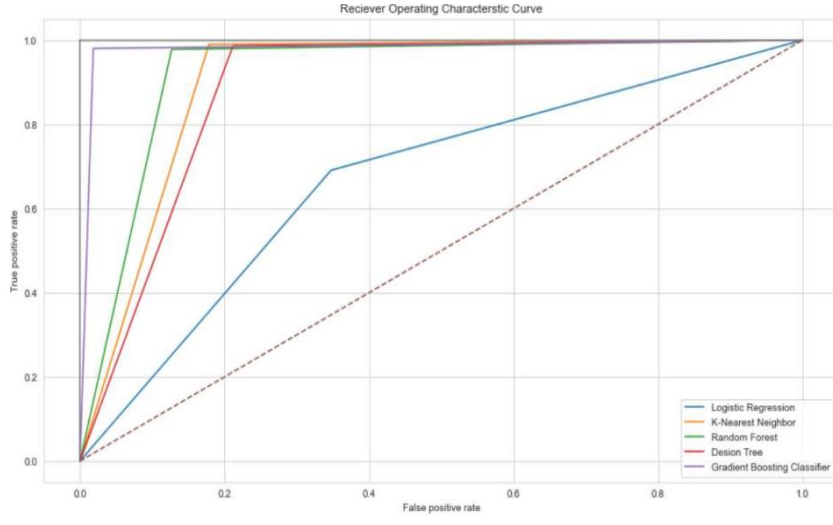
For each analysis, performance indicators, model training, computational time, and performance analysis score are supplied. The created models results are visualized using charts that illustrate training and validation accuracy. Metrics including as precision, recall, and F1-Score were used to demonstrate test accuracy. In tabular style, a comparison of the accuracy of the several created models. Hyper-parameters that were employed during model training were also mentioned.

The performance all the models and to decide which model is best to predict CHD over the given dataset or any further research required to gather more insights. Below chart in Table 5 shows different model accuracy. Finally, it shows Gradient boosting is the best model after hyper parameter tuning.

Table 5: Model Accuracy

	Model	Accuracy
0	Logistic Regression	67.170228
1	K-Nearest Neighbour	90.567428
2	Random Forest	92.557111
3	Decision Tree	88.725129
4	Gradient Boosting	98.084009

Figure 7: ROC different Models



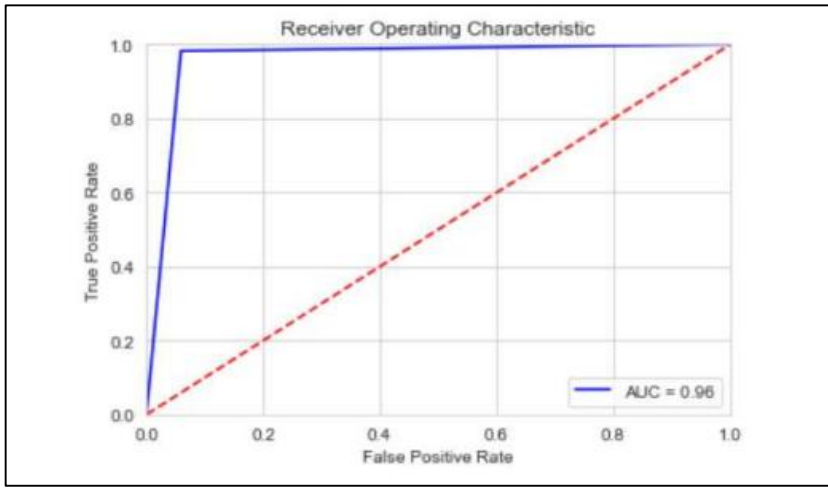
The novel approach commenced computes prediction accuracy. Gradient Boosting classifier after hyperparameter tuning seems to overfitted and losing generalization. Model's ability is to generalise is critical to its success. Even if it can make good predictions for the training data, it will make false predictions when given fresh data, rendering the model useless. As an ensemble, a stacking strategy is employed to boost the model's accuracy. It's a combination of three classifiers: Random Forest, Gradient Boosting, and K-nearest neighbour, all of which were tuned after hyperparameter tweaking.

The unique approach that has never been attempted before. The accuracy of Stacking CV Classifier is 96.16% and AUC is 0.96 as shown in Figure 8.

Figure 8: ROC Stacking CV Classifier ensemble

Accuracy of StackingCVClassifier: 96.16801768607222

	precision	recall	f1-score	support
0	0.98	0.94	0.96	1356
1	0.94	0.98	0.96	1358
accuracy			0.96	2714
macro avg	0.96	0.96	0.96	2714
weighted avg	0.96	0.96	0.96	2714



The gradient boosting model, followed by Random Forest and KNN models, produced the highest AUC score using the Framingham CHD dataset. The precision accuracy is determined by the properties that a data set comprises and the algorithm supplied; the best algorithms that provide precision are chosen to determine the forecast. The simulation models are built around ten functions, and the modelling methodologies' precision is determined. Tuning the features and selecting the appropriate feature the models were run with all the attributes the accuracy was of 85-90 percent. The features were reduced, and the accuracy remained the same. With progress attributes were removed one by one until the point is reached where the accuracy dropped to 80-85 percent. The conclusion is that for higher accuracy, there should be a minimum of 10 attributes. The chosen attributes are reduced, accuracy suffers, otherwise, performance suffers.

In comparison to previous relevant articles, various techniques for heart disease were established utilizing the same data sets. However, the methodologies for categorizing and forecasting cardiac disease were shown to be compatible with the existing models. The proposed grouping of learning ensembles and simulation models has the potential in reducing the number of inaccurate diagnoses and medical treatments that have harmed patients' health. The suggested community classification and prediction models help enhance the chances of survival and save millions of lives for heart disease patients by allowing early and precise detection of heart disease. This experiment shows that ensembles outperform single classifiers, with stacked ensembles providing superior accuracy.

Additional classification techniques, such as support vector machine (SVM), Nave Bayes, and several forms of artificial neural networks, are applied to improve the research. By comparing the results of different classification algorithms, data scientists can discover new options for assisting the healthcare industry in the fight against the world's deadliest diseases. It will also be useful to diversify the datasets even more. While the datasets used in this study varied in many respects, they were all limited to binary replies based on the demographical, behavioral, and medical data of the patients. The adaptability of the applications could be improved by expanding the research effort to explore heart picture data for disease prediction and detection using computer vision.

Apart from the standard regressors and classifiers, the ensemble has the highest number of accuracies, followed by Random Forest classifier, and finally KNN. Gradient Boosting operates admirably and provides excellent accuracy in this study, as does the stacking classifier. However, because Gradient Boosting is so close to perfect in accuracy due to oversampling may cause generalisation problems that prevent it from

performing from other algorithms in real-world data. The novel approach of Stacking CV classifier ensemble of Random Forest, Gradient Boosting, and KNN will help to mitigate this issue.

6.0 Conclusion

Various common machine learning algorithms have been presented in this research, basic functioning mechanisms, and applied to real-world datasets, with a study being conducted to discover the best classifier. Finally, different binary classification approaches are examined for predicting heart disorders or heart-related diseases. Gradient Boosting has been demonstrated to be the most accurate classification method for predicting the risk of heart disease, with a 98 percent accuracy rate but overfitting of the training dataset is a concern. The Stacking CV classifier ensemble proposed model has substantially good accuracy of 96.16%

Finally, the most significant factor in the world of healthcare is the question of why. The reason why certain algorithms function better than others on certain ailments or diseases is far more important than the performance metrics of the algorithms. This boils down to debating the algorithms' interpretability, a thorough comprehension of their structures, decision factors, and any underlying relationships. For this critical next step to occur, the research must bring together experts and resources from the healthcare domain to scrutinize the analytical methods and results from a new, but equally important, perspective. This partnership necessitates excellent communication between the two domains, as well as, if necessary, the establishment of training methods for medical organizations to completely appreciate the results of machine learning models to properly interpret them.

References

[1] Alanazi, R. (2022). Identification and prediction of chronic diseases using Machine Learning approach. *Journal of Healthcare Engineering, 6,* 1-9. Retrieved from DOI:10.1155/2022/2826127

[2] Alex, M. & Shaji, S. (2019). Prediction and diagnosis of heart disease patients using data mining technique. *International Conference on Communication and Signal Processing*. Retrieved from DOI:10.1109/ICCSP.2019.8697977

[3] Anbuselvan, P. (2020). Heart disease prediction using machine learning techniques. *International Journal of Engineering Research and Technology, 9*(11), 515-518.

[4] Asif, S., Wenhui, Y., Jinhai, S. & Jin, H. (2021). An ensemble machine learning method for the predicton of heart disease. *4th International Conference on Artficial Intelligence and Big Data*. Retrieved from 10.1109/ICAIBD51990.2021.9459010

[5] Bharti, R., Khamparia, A., Shabaz, M. & Dhiman, G. (2021). Predcition of heart disease using a combination of machine learning and deep learning. *Computational Intelligence and Neuroscience,* 1687-5273. Retrieved from DOI:10.1155/2021/8387680

[6] Cardiovascular diseases (2022 Feb). Retrieved from https://www.who.int/ healthtopics/cardiovasculardiseases.

[7] Chen, C. & Zhang, X. (2021). Early warning methods of epidemiological risks based on data mining. *International Conference on High Performance Big Data & Intelligent Systems*. Retrieved from DOI:10.1109/HPBDIS53214.2021.9658444

[8] Kanchan, B. D. & Mahale, K. M. (2017). Study of machine learning algorithms for special disease prediction using principal of component analysis. *International Conference on Global Trends in Signal Processing, Information computing and communication*. Retrieved from 10.1109/ICGTSPICC.2016.7955260

[9] Dinesh, K. G., Arumugaraj, K., Santhosh, K. D. & Mareeswari, V (2018). Prediction of cardiovascular disease using machine learning algorithms. *International Conference on Current Trends towards Converging Technologies*. Retrieved from DOI:10.1109/ICCTCT.2018.8550857

[10] Dwivedi, A. K. (2018). Performance evaluation of different machine learning techniques for prediction of heart disease. *Neural Computing and Applications, 29*(10), 685-693.

[11] Galetsi, P., Katsaliaki, K. & Kumar, S. (2019). Values, challenges and future directions of big data analytics in healthcare: A systematic review. *Social Science and Medicine, 241*(5), 112533.

[12] Galetsi, P. & Katsaliaki, K. (2019). A review of the literature on big data analytics in healthcare. *Journal of the Operational Research Society, 71*(1), 1-19.

[13] Gao, X., Ali, A. A., Shaban, H. & Anwar, E. M. (2021). Improving the accuracy for analyzing heart diseases prediction based on the ensemble method. *Complexity,* 1-10. Retrieved from DOI:10.1155/2021/6663455

[14] Karim, A., Jonkman, M., Hasan, M. Z. & Ghosh, P. (2021). Use of efficient machine learning techniques in the identification of patients with heart diseases. *ACM International Conference Proceeding Series*. Retrieved from DOI: 10.1145/3471287.3471297

[15] Ghosh, P., Azam, S., Jonkman, M. & Karim, A. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. *IEEE Access*. Retrieved from DOI: 10.1109/ACCESS.2021.3053759

[16] Goel, S., Deep, A., Srivastava, S. & Tripathi, A. (2019). Comparative analysis of various techniques for heart disease prediction. *4th International Conference on Information Systems and Computer Networks (ISCON)*. Retrieved from DOI:10.1109/ISCON47742.2019.9036290

[17] Huang, H., Tan, J. & Hua, D. (2021). Data mining of association between hyperuricemia and common chronic diseases based on evolutionary apriori algorithm (EAA). *IEEE 6th International Conference on Cloud Computing and Big Data Analytics*. Retrieved from DOI:10.1109/ICCCBDA51879.2021.9442490

[18] Krishnan, S. J. & Geetha, S. (2019). Prediction of heart disease using machine learning algorithms. *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*. Retrieved from DOI:10.1109/ICIICT1.2019.8741465

[19] Modepalli, K., Gnaneswar, G., Dinesh, R. & Sai, Y. R. (2021). Heart disease prediction using hybrid machine Learning model. *6th International Conference on Inventive Computation Technologies (ICICT)*. Retrieved from DOI:10.1109/ICICT50816.2021.9358597

[20] Kohli, P. S. & Arora, S. (2018). Application of machine learning in disease prediction. *4th International Conference on Computing Communication and Automation (ICCCA)*. Retrieved from DOI: 10.1109/CCAA.2018.8777449

[21] Krishnani, D. (2019). Prediction of coronary heart disease using supervised machine learning algorithms. *IEEE Region 10 Conference (TENCON)*. Retrieved from DOI: 10.1109/TENCON.2019.8929434

[22] Kwakye, K. & Dadzie, E. (2021). Machine learning-based classification algorithms for the prediction of coronary heart diseases. *Retrieved from* https://www.researchgate.net/publication/356746472_Machine_Learning-Based _Classification_Algorithms_for_the_Prediction_of_Coronary_Heart_Diseases

[23] Mutyala, N. K., Koushik, K. V & Krishna, K. D. (2018). Prediction of heart diseases using data mining and machine learning algorithms and tools. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*. Retrieved from DOI:10.13140/RG.2.2.28488.83203

[24] Li, J., Haq, A., Swati, S. & Khan, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. *IEEE Acsess,* 107562–107582. Retrieved from DOI: 10.1109/ACCESS.2020.3001149

[25] Pavithra, V. & Jayalakshmi, V. (2021). Comparative study of machine learning classification techniques to predict the cardiovascular diseases using HRFLC. *5th International Conference on Intelligent Computing and Control Systems (ICICCS)*. Retrieved from DOI:10.1109/ICICCS51141.2021.9432105

[26] Qayyum, A., Qadir, J., Bilal, M. & Al-Fuqaha, A. (2021). Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering, 14,* 156-180.

[27] Anirban, D. (2021). Prediction of coronary heart diseases using machine learning-based. *Liverpool John Moores University*.

[28] Sivabalaselvamani, D., Selvakarthi, D., Loganathan, R. & Eswari, S. N. (2021). Investigation on heart disease using machine learning algorithms. *International Conference on Computer Communication and Informatics (ICCCI)*. Retrieved from DOI:10.1109/ICCCI50826.2021.9402390

[29] Rajdhan, A., Agarwal, A., Sai, M. & Ghuli, P. (2020). Heart disease prediction using machine learning. *International Journal of Engineering Research & Technology (IJERT), 9*(4), 659-662.

[30] Rosengren, A., Smyth, A., … Yusuf. S. (2019). Socioeconomic status and risk of cardiovascular disease in 20 low-income, middle-income, and high-income countries: The Prospective Urban Rural Epidemiologic (PURE) study. *The Lancet Global Health, 7*(6), e748–e760.

[31] Sharma, H. & Rizvi, M. A. (2017). Prediction of heart disease using machine learning algorithms: A survey. *International Journal on Recent and Innovation Trends in Computing and Communication, 5*(8), 99-104.