IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 26, Issue 3, Ser. 1 (May. – June. 2024), PP 55-59
www.iosrjournals.org

Artificial Intelligence In Cardiac Health: Predictive Modeling Of Heart Disease

Dr. Vrunda Kusanur[1], Dr. Sujaya B L[2], Dr. Rashmi S Bhaskar[3]
[1,2,3] *(Ece, Bnm Institute Of Technology, Bangalore, Karnataka / Vtu, India)*

Abstract:
In today's world, a precise method for predicting cardiac disease is essential with emotionally supportive system. To determine if a patient has a cardiac illness or is in a normal condition, data mining techniques are frequently used. In this work, a prediction model for forecasting cardiac diseases is suggested using the Naive Bayes (NB) and Random Forest algorithms. Extensive testing on a large-scale coronary health dataset validates the model's better performance compared to current methods, which is primarily reflected in enhanced prediction accuracy.
Key Word*: Data Mining, Naive Bayes, Random Forest*

Date Of Submission: 29-04-2024 Date Of Acceptance: 09-05-2024

I. Introduction

Machine learning is a rapidly expanding field with several research prospects. Machine learning is an evolving technology that allows computers to learn automatically from previous data. Machine learning employs a variety of algorithms to make predictions based on previously collected data or information. Machine learning is being used in a wide range of fields, including traffic prediction, medical diagnosis, email filtering, voice identification, picture and image recognition, and marine wildlife preservation.

According to World Health Organisation (WHO), heart disease kills approximately 12 million people globally each year. Healthcare machine learning aids in analyzing extensive and complex medical data, enabling the extraction of valuable clinical insights. Subsequently, doctors can utilize this information to sustain and improve medical care delivery.

Machine learning uses past data to create predictions. Machine learning models learn from both data and experience throughout the training phase of the algorithms. Despite ongoing challenges from past decades in predicting diseases based on patient symptoms and history, machine learning algorithms offer promise in successfully addressing healthcare challenges. Using machine learning models, the system that effectively cleans and process the data would be created to produce quick results. This paper enables healthcare professionals to make well-informed decisions about patient diagnoses, leading to improved treatment selections and ultimately elevating the quality of healthcare services for patients.

The main goal for conducting this study is to suggest a model to forecast the onset of cardiac disease. Furthermore, the purpose of this work is to identify the best classification scheme for diagnosing heart illness in a patient. For this purpose, three classification models such as Naive Bayes, Decision Tree and Random Forest are compared at multiple levels of assessment. The prediction of heart illness is a critical task and demands the highest level of accuracy from the machine learning algorithms. As a result, the three algorithms are assessed using a range of levels and evaluation technique types. It attempts to assess whether a patient is at risk of developing cardiovascular disease. The results of the heart disease prediction demonstrate the highest level of accuracy. This will help researchers and doctors to better understand the problem and to determine the best strategy to identify heart illness.

The main objectives of this work include the simplification of diagnosing process of heart disease, to improve the efficiency of diagnostics using machine learning, to forecast cardiac disease and take preventive measures and to guarantee that the medical diagnosis is available to all people.

II. Related Work

A unique strategy for identifying key information using machine learning techniques, which improved the accuracy of cardiovascular disease prediction is suggested in this work.[1]. The forecast model is presented by combining various features and classification approaches. The prediction model for heart disease with the hybrid random forest with a linear model (HRFLM) results in improved performance with an accuracy level of 88.7%.

An innovative model-creation strategy is recommended for handling real-world issues.[2]. The algorithms are validated using a 5-fold cross-validation method. The investigation reveals that the Extreme Gradient Boosting

Classifier with GridSearchCV achieves the greatest and almost equivalent testing and training accuracies of 100% and 99.03% for both datasets (Hungary, Switzerland, and Long Beach V, and UCI Kaggle). Furthermore, the study reveals that the XGBoost Classifier without GridSearchCV achieves the greatest and almost equal testing and training accuracies of 98.05% and 100% for both datasets (Hungary, Switzerland & Long Beach V and UCI Kaggle). In addition, the analytical results of the suggested approach are compared to current heart disease prediction research. It is clear that among the suggested approaches, the Extreme Gradient Boosting Classifier with GridSearchCV produces the best hyperparameter for testing accuracy.

Machine Learning (ML) algorithms is used to forecast CHD using historical medical data. This research focuses on employing three supervised learning techniques: Naive Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT) to identify correlations in CHD data and improve prediction rates. This paper focuses on employing three supervised learning techniques: Naive Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT) to identify correlations in CHD data and improve prediction rates. The ML models are trained using the South African Heart Disease dataset of 462 instances and validated via 10-fold cross validation. Experiential results using several performance evaluation metrics show that probabilistic models produced by NB are promising for identifying CHD.[3].

Proposed a new hybrid classifiers, such as the Decision Tree Bagging Method (DTBM), Random Forest Bagging Method (RFBM), KNearest Neighbours Bagging Method (KNNBM), AdaBoost Boosting Method (ABBM), and Gradient Boosting Boosting Method (GBBM) by combining traditional classifiers with bagging and boosting methods during the training process. In this study, to forecast cardiac disease, appropriate features are chosen by using the Relief, and Least Absolute Shrinkage and Selection Operator (LASSO) techniques. In this work, the Accuracy (ACC), Sensitivity (SEN), Error Rate, Precision (PRE) and F1 Score (F1) of ML model, along with the Negative Predictive Value (NPR), False Positive Rate (FPR), and False Negative Rate (FNR) are computed and compared for all the ML algorithms. This work concludes that the proposed model generated the maximum accuracy of 99.05% when RFBM and Relief feature selection methods are employed.[4].

A technique for easily distinguishing and classifying people with heart disease from healthy ones is devised.[5]. This paper discussed all the classifiers, feature selection techniques, pre-processing methods, validation methods, and classifier performance assessment measures employed in this study. Furthermore, Receiver Optimistic Curves and Area Under Curves for each classifier were calculated. The proposed system's performance has been validated across all features as well as a subset of them. The reduced number of features has an influence on classifier performance in terms of accuracy and duration of execution.

III. Design Methodology

Figure 1 presents the design methodology of the heart disease prediction system. The following steps are involved in the prediction of heart disease using Machine Learning models.

Input: Dataset of heart illness.
Output: Indicate if a person has cardiac disease or is healthy.
Step 1: The heart data set is loaded.
Step 2: Apply preprocessing filter discretization and InterQuartile Range (IQR).
Step 3: Divide the datasets into training and test sets.
Step 4: The heart disease data set is used to train the models for Random forest and Naive Bayes algorithms.
Step 5: Evaluate the Random Forest and Naive Bayes algorithms' accuracy.
Step 6: Choose the classifier that has the highest accuracy.
Step 7: Predict the illness using the input data and the classifier's output.

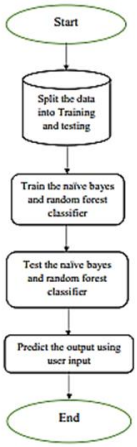

Figure 1: Design Methodology of Heart Disease Prediction Model

IV. Proposed System

This work employs Naive Bayes and Random Forest algorithms for heart illness prediction.

Naive Bayes

Naive Bayes is a machine learning technique that uses probability to categorise data as depicted in Figure 2. The Naive Bayes classifier is the most basic Bayesian network classifier that makes use of the Bayes hypothesis and the string independence of attributes assumption. Naive Bayes is a simple approach for constructing classifiers. This model assigns class labels to problem cases, where the class labels are selected from a predetermined set. Class labels are denoted as vectors of feature values. Naive Bayes has the benefit of requiring minimal training data to accurately estimate classification parameters.

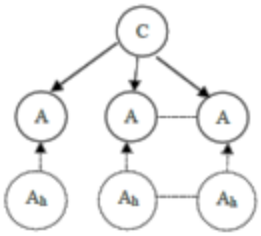

Figure 2: Naive Bayes Classifier

Naive Bayes classifier is implemented as follows.
Input: A set of database
Output: Naive Bayes classifier
Step 1: For each value of c of class C
Step 2: Estimate probabilities P(C) from Database D
Step 3: For attributes Ai and Aj
Step 4: Calculate P (ai|aj, c) from D
Step 5: Evaluate conditional mutual information MI=IP(Ai; Aj| C) and weights Wij from D.

Random Forest

Random forest is an ensemble learning approach for classification, regression, and other tasks that involves the construction of a large number of decision trees during training as shown in Figure 3. For classification problems, the random forest produces the output as the class chosen by the majority of trees. For regression problems, the mean or average prediction from each tree is returned. The practice of decision trees to overfit to their training data set is improved for by random decision forests. Although Random Forest are less accurate than Gradient Boosted Trees, they perform better than choice trees in most cases. The performance of Random Forest, however, may be influenced by the features of the data.

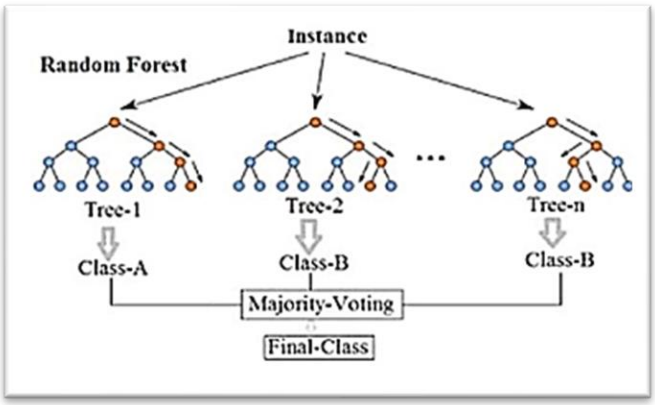

Figure 3: Naive Bayes Classifier

Random Forest classifier is implemented as follows.
Input: A set of database
Output: Random Forest classifier
Step 1: In Random forest *n* number of random records are selected from the *k* number of records in the data set.
Step 2: For each sample, separate decision trees are created.
Step 3: The separate output is produced by each decision tree.

Step 4: Final output is measured using Majority Voting or Averaging for classification and regression respectively.

Details of Dataset

The large amount of data has been used from Kaggle. There are 14 features, namely, *age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca , thal and target* that are being used for forecasting heart illnesses as shown in the Figure 4.

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
58	0	0	100	248	0	0	122	0	1	1	0	2	1
58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
71	0	0	112	149	0	1	125	0	1.6	1	0	2	1
43	0	0	132	341	1	0	136	1	3	1	0	3	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
52	1	0	128	204	1	1	156	1	1	1	0	0	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
54	1	0	124	266	0	0	109	1	2.2	1	1	3	0
50	0	1	120	244	0	1	162	0	1.1	2	0	2	1
58	1	2	140	211	1	0	165	0	0	2	0	2	1
60	1	2	140	185	0	0	155	0	3	1	0	2	0
67	0	0	106	223	0	1	142	0	0.3	2	2	2	1
45	1	0	104	208	0	0	148	1	3	1	0	2	1

Figure 4: Dataset Details

The details of input dataset features are illustrated as follows.

- *Age* : Age in years
- *Sex* (value 1: Male; value 0 : Female)
- *CP*(Chest pain type):

value 1: typical type 1 angina
value 2: typical type angina
value 3: non-angina pain
value 4: asymptomatic

- *Trestbps:* The individual's resting blood pressure (mm Hg on admission to the hospital)
- *chol:* The individual's cholesterol measurement in mg/dl. Numeric value(140mm/Hg).
- *Fbs* (Fasting Blood Sugar): The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false).
- *restecg*: resting electrocardiographic results;

Value 0: showing probable or definite left ventriular hypertrophy by Estes' criteria
Value 1: normal
Value 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV).

- *thalach*: The individual's maximum heart rate reached.
- *exang*: Exercise induced angina (1 = yes; 0 = no)
- oldpeak: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot.
- *slope*: the slope of the peak exercise ST segment — 0: downsloping; 1: flat; 2: upsloping
- *ca*: The number of major vessels (0–3)
- *thal*: A blood disorder called thalassemia and its values are detailed in Table no1.

Table no1: Description of thal values

thal Value	Description
0	NULL
1	fixed defect (no blood flow in some part of the heart)
2	normal blood flow
3	reversible defect (a blood flow is observed but it is not normal)

target: The "target" field denotes to the existence of heart disease in the patient. '0' indicates no disease and value '1' indicates disease.

The correlation method is used to evaluate the importance of these attributes in the estimation of heart disease. The results illustrate that only ten characteristics—*age, sex, cp, restecg, thalach, exang, oldpeak, slope, ca, and thal*—are determined to be the most significant in calculating heart illnesses.

V. Results And Discussions

The results of Naive Bayes and Random Forest classifiers are used to predict the heart disease. From Figure 5 to Figure 8 depict the results of proposed prediction model for different values of *age, sex, cp, restecg, thalach, exang, oldpeak, slope, ca, and thal.*

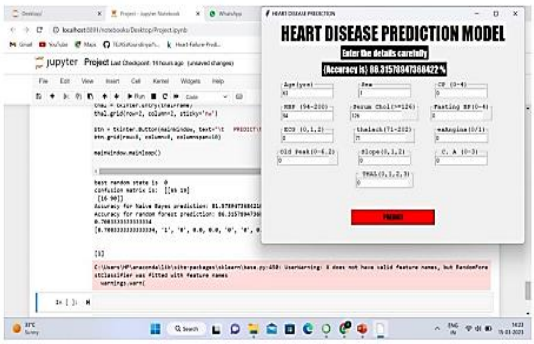

Figure 5: Interface with low values

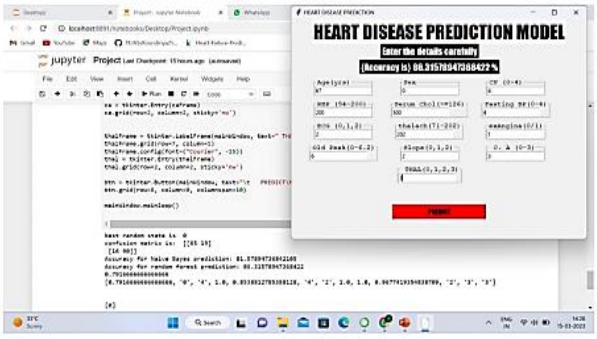

Figure 6: Interface with high values

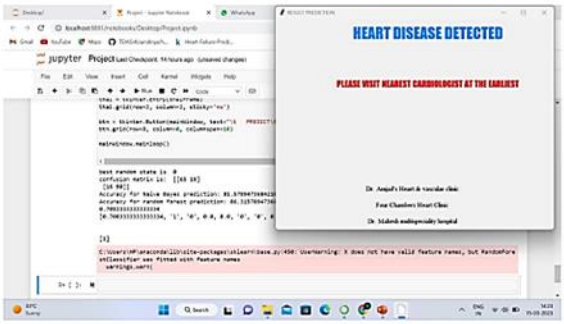

Figure 7: Heart disease detected

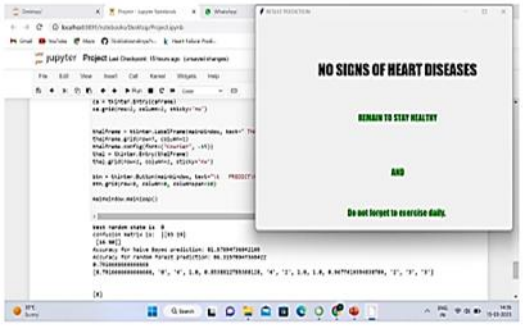

Figure 8: No sign of heart disease detected

VI. Conclusion

Proposed work used Naive Bayes classification and the Random Forest algorithms to predict the occurrence of heart disease. This experimental setup is designed to develop a prediction system, investigate, and forecast the risk of cardiovascular sickness, as heart disease is the leading cause of death globally. It is feasible to avoid death at an early stage using this model. The proposed model is conceptually believed to be a promising method for clinical data sets such as heart disease with dependent features for disease identification based on its performance results.