Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques

Dr.Abdul Khadeer
Associate Professor
Department of CSE
Deccan College of Engineering and Technology
Affiliated to Osmania University
abdulkhadeer@deccancollege.ac.in

Mohammed Qayamuddin
PG Scholar
Department of CSE
Deccan College of Engineering and Technology
Affiliated to Osmania University
mohammedqayamuddin725@gmail.com

Abstract: A major cause of mortality in the world is cardiovascular illness, and early detection is critical in reducing the rates of mortality. Cardiac disease is a complex medical condition, and its exact prognosis cannot be made because of the absence of the continuous monitoring opportunities. Based on the dataset of Heart Disease, most feature selection methods, like ANOVA F-statistic (ANOVA FS), Chi-squared test (Chi2 FS), and Mutual Information (MI FS), were applied to determine the relevant predictors. Synthetic Minority Oversampling Technique (SMOTE) was applied to address data imbalance to enhance the efficacy of models. In-depth classification methodology was used using different machine learning models and ensemble techniques. A Stacking Classifier which combines Boosted Decision Trees, Extra Trees and LightGBM has scored very high with 100 percent accuracy in all the feature selection approaches. The high performance highlights the effectiveness of the advanced ensemble learning to generate plausible heart disease predictions, and there are possibilities of using strong feature selection with advanced classification models to successfully analyze medical data. The approach represents the capacity to promote early diagnosis and patient outcomes.

"Index Terms - *Cardiovascular disease, heart disease, machine learning app, ML algorithms, SDG 3, SHAP, SMOTE.".*

1. INTRODUCTION

Heart disease is one of the leading causes of death throughout the world with cardiovascular diseases (CVD) being one of the leading mortal causes of death across the world. The heart is a muscular organ that is very vital to the circulatory system since it pumps blood all over the body. This is a complicated system that includes arteries, veins and capillaries which carry oxygen and nutrients to organs and tissues. Alterations in the progress of the normal blood flow cause various kinds of heart illness due to which it is usually referred to as cardiovascular diseases (CVD). According to the World Health Organization (WHO) an average of more than 17.5 million deaths is as a result of heart disease and stroke with more than 75 percent of the deaths occurring in low- and middle-income countries. This alarming fact highlights the growing societal health risk posed by heart disease on the global level with cardiovascular disease related deaths (cardiac attacks and strokes) comprising 80% of all cardiovascular disease related mortality [1].

The spread of cardiovascular diseases has led to a global focus on early detection, prevention, or treatment protocols. In line with the United Nations sustainable development goal 3 that emphasizes the importance of health and well-being, addressing the problem of cardiovascular illnesses has become one of the priorities when it comes to improving the health outcomes in the world. Some of the common risk factors that contribute to cardiovascular disease include tobacco consumption, old age, family history, high cholesterol level, sedentary living, high blood pressure, obesity, diabetes, and psychological stress. Lifestyle changes such as quit smoking, regular exercises, keeping body weight. and relieving stress have been identified to reduce the risk of heart diseases [2]. Diagnostic tools including electrocardiograms (ECGs), echocardiograms, cardiac MRIs and blood tests are also commonly used in conjunction with lifestyle changes to diagnose heart disease. In certain cases, medical procedures such as as angioplasty, coronary artery bypass grafting and implantation of mechanical devices such as pacemakers and defibrillators can be required to treat [3].

Healthcare technology innovations, notably through Big Data and Electronic Health Records (EHRs), have made it possible to use a large amount of patient data to make predictions. The current use of ML techniques to analyze large quantities of healthcare systems data is increasingly utilized to draw meaningful conclusions about the likelihood of heart disease. By examining and utilizing patient data of various patient groups, risk factors, and patient outcomes in the form of a diagnosis, ML can help medical professionals establish high-risk patients and enable early intervention. The healthcare industry is being transformed by this kind of methodology into a more accurate and efficient means of diagnostic, predictive and customized treatment plans [4][5].

2. RELATED WORK

The cardiovascular diseases (CVD) including coronary heart disease remain a health problem in the world, giving rise to a significant proportion of deaths in the world. The growth of healthcare data and advances in ML methods prompted the growth in the effort to predict and diagnose cardiac disease with greater accuracy. The ability to analyze a large amount of data using ML creates new possibilities of risk factors, forecasting, and improving an early diagnosis. The literature review focuses on the various studies that used different models and techniques of ML to forecast heart disease based on their findings.

In their study, Yang et al. [6] selected the use of ML to determine the risk factors of coronary heart disease. They focused on large data analysis in their research, and their study demonstrated how ML models (such as DT, RF, support vector machines (SVM)) could identify significant risk factors in patient data well. The study found the best risk factors to predict heart disease to be high cholesterol, age, and familial history. This paper has highlighted the importance of data preparation, feature selection and the need to have a sound dataset in improving model performance. The results confirmed that ML can be effectively used as a tool of early cardiac disease detection, especially when combined with large and comprehensive datasets.

Ngufor et al. [7], also, compared different ML tools in the prediction of cardiac disease. The authors have made a comparative analysis of some of the common methods like SVM, DT, KNN and artificial neural networks (ANNs). Their results showed that ensemble methods, such as bagging and boosting gave better predictive accuracy than univariate models. The paper has emphasized the role of feature selection since irrelevant features may reduce the accuracy of the model. As this review revealed, it is possible to find quite a number of algorithms that predict cardiac disease, but the choice of strategy to use is heavily conditioned by the characteristics of the data, the computing capabilities at hand, and the specific requirements of the prediction task.

Farag et al. [8] focused on the project of improving the prediction of heart diseases by using the boosting and bagging methodologies. Algorithms like AdaBoost and bagging algorithms like the RF were tested on their effectiveness to enhance the accuracy of prediction. The study revealed that ensemble was more robust compared to individual classifiers in terms of reduction of variance and stability of prediction. The research suggested that, the overfitting problem that is usually encountered when modeling heart disease predictions, could be relieved through the incorporation of boosting and bagging. This paper has found the importance of using a plethora of classifiers simultaneously to achieve the best performance.

Zhang et al. [9] explored the use of XGBoost which is a gradient boosting algorithm in the clinical prediction of coronary heart disease. Their study showed that XGBoost was much better than traditional methods like logistic regression and SVM in terms of accuracy and interpretation. This ability of XGBoost to deal with unbalanced data, which is frequently a problem with heart disease prediction, makes it particularly suitable in medical data. The study showed that hyperparameter optimization was necessary to maximize the performance of the model. Such remarkability is possible because XGBoost performs better clinically because it is efficient, scalable, and able to provide plausible predictions with minimum chances of overfitting.

Liu et al. [10] conducted comparative study of different ML methods, such as DT, support vectors machines, and RF applied in prediction of heart disease. Their study demonstrated that the SVM that applied the radial basis function kernels had the highest prediction accuracy in comparison with the reviewed algorithms. However, they noted that DT and RF were better in terms of interpretability, which is critical in a medical setting. The researchers concluded that SVM was the most accurate, but the choice of an algorithm is to be made in terms of accuracy and interpretability, especially when the model is supposed to be used by medical practitioners to make decisions.

Hussein et al. [11] compared the different ML techniques to diagnose cardiac disease, which include, the KNN, DT, and the artificial neural networks. The aim of the study was to use the models to determine their diagnostic performance using a sample of patient health records. Their results showed that KNN and DT had high performance with less computing requirements making them suitable in real-time use in a clinical setting. Even though artificial neural networks were more accurate, they required more computer resources and were not easy to interpret. The paper has emphasized that efficacy and resource constraints should be considered when implementing ML models in healthcare systems.

Akbar et al. [12] did a thorough evaluation of different ML techniques in cardiac disease prediction. They evaluated such methodologies as DT, KNN, support vectors machine, RF, and neural networks and provided an overview of its pros and cons in predicting heart disease. The researchers concluded that the ensemble methods, in particular, the RF, achieved the highest accuracy because of the ability to reduce overfitting and handle noisy data. However, the analysis highlighted the challenge of determining the most relevant features using large volumes of data as the selection of features is essential to the effectiveness of the model. The paper has highlighted the importance of dealing with missing data and making the training sample representative and balanced.

The study by Zarshenas et al. [13] was a comparative study on ML strategies in predicting heart diseases, with the highlight being on SVM, DT, KNN, and logistic regression. They have found out that both SVM and RF were the best in predictive performance, though SVM was more accurate under certain cases. They indicated how the preprocessing methods such as normalizing and scaling of features contribute to the effectiveness of ML models. The results have shown that hybrid models that combine the strengths of most methods can be the possible way to develop new studies in predicting cardiac diseases.

The issue of feature selection and data preprocessing as factors of enhancing the effectiveness of heart disease prediction models was another common theme in the reviewed papers. A variety of studies indicate that ensemble strategies, including bagging and boosting, are effective to improve the accuracy of the models, whereas some of them suggest that hybrid models, consisting of using different ML algorithms, can yield better results. Furthermore, despite the fact that other traditional models, such as DT and logistic regression, are still popular, recent studies report that advanced algorithms, such as XGBoost and neural networks, can be more effective than the current methods, particularly when using complex and high-dimensional data.

The increased availability of healthcare data and the advent of ML techniques have provided novel possibilities in the early detection and diagnosis of cardiac disease. The ML models help healthcare providers to spot at-risk persons and provide timely intervention and reduce the load of cardiac disease. However, the problems such as data quality, feature selection, model interpretability, and computational efficiency should be addressed so that the full capabilities of these technologies can be applied to clinical practice. A combination of advanced ML algorithms and specific knowledge of healthcare specialists is likely to drive the new achievements in predicting heart diseases.

3. MATERIALS AND METHODS

The proposed solution would have developed an accurate cardiac diseases predictive model based on ML and advanced ensemble techniques. Preprocessing of Heart Disease dataset is performed, and important features are determined using ANOVA F-statistic, Chi-squared test and Mutual Information methods. Synthetic Minority Oversampling Technique (SMOTE) is used to correct the imbalance between classes, therefore, obtaining the balance in data distribution.

The system evaluates numerous ML models, which are NB, SVM, XGBoost, AdaBoost, Bagging Classifier, DT, KNN, RF, and Logistic Regression. Voting Classifier This combines the outputs of more than one model to produce a more accurate and resilient output. Stacking Classifier merges Boosted DT, ET and LightGBM with the goal to use the synergistic benefits. This integrative approach aims to enhance the precision and reliability of cardiac disease prediction so as to enable the detection and better clinical decisions.

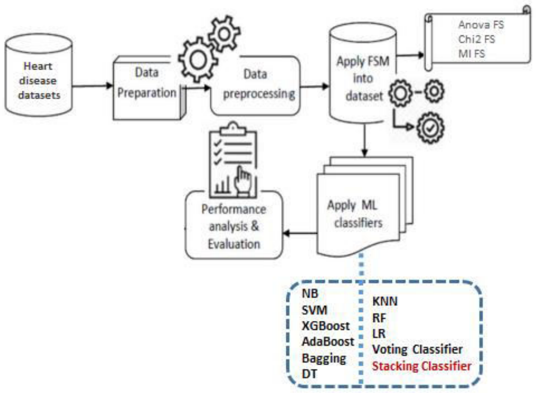
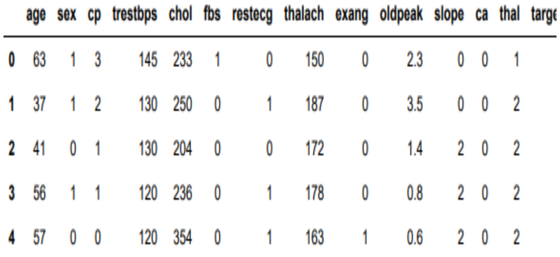


Fig.1 Proposed Architecture

The following graphic (Fig. 1) shows a flowchart of a heart disease prediction model. The approach starts with the preparation and pre-processing of heart disease data. Then, feature selection (ANOVA FS, Chi-squared FS, MIFS) are used. The information is then fed to a series of ML classifiers (NB, KNN, SVM, RF, XGBoost, LR, AdaBoost, Voting Classifier, Bagging, Stacking Classifier, DT). Performance analysis and testing of the model is performed to identify how accurate and effective it is as a predictor of cardiac disease.

i) Dataset Collection:

The dataset that is used to predict the heart disease includes 303 samples and 14 features including both numerical and categorical variables. These characteristics include such key details of the patient as age, sex, type of chest pain (cp), resting blood pressure (trestbps), cholesterol levels (chol), fasting blood glucose (fbs), electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), and ST segment depression induced by exercise (oldpeak), slope of the peak exercise ST segment (slope), number of major vessels visualized by fluoroscopy (ca), and thalassemia (thal). Goal variable is a categorical variable that relates to the presence or absence of cardiac disease. After the application of feature selection techniques, such as ANOVA F-statistic, Chi-squared test, and Mutual Information, different sets of features were determined to improve the accuracy of the model and their efficiency, which maximizes the dataset to predict the outcomes of heart diseases.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	targe
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	

Fig.2 Dataset Collection Table – Heart Disease Data

ii) Pre-Processing:

The pre-processing is a critical stage in the ML dataset preparation process. It involves cleaning up and remodeling of data to ensure accuracy, effectiveness, and relevance. Splendid control of the absent values, encoding and feature choice largely enhances the model efficacies and resilience.

a) Data Processing: The first step in the process of data processing involves cleansing, which involves removing the missing values and correcting the discrepancies. The data is optimized by removing unnecessary columns. Categorical variables are encoded with the help of labeling and they are separated into input (X) and output (y) datasets to ensure the proper structuring to analyze them. Such steps ensure that the data is ready to train a model.

b) Data Visualization: Data visualization aids in the understanding of the relationship between variables and the uncovering of hidden patterns. A correlation table is compiled to test the interdependence between the numerical characteristics and sample results are plotted to investigate the distribution of data and trends. It helps to determine the relevant characteristics and explain their effect on the target variable.

c) Label Encoding: Label encoding is used to encode categorical labels using numerical values so that the models can deal with non-numeric data. It is a process that converts every category into a specific integer making them suitable to ML algorithm types that require numbers. The label encoding would be particularly beneficial when categorical data has a natural order.

d) OverSampling: SMOTE (Synthetic Minority Over-sampling Technique) is used to correct for the imbalance in classes by creating artificial samples of the minority group. This technique supports the formation of a balanced data through oversampling of the underrepresented group, thereby avoiding the possibility of the model being biased against the majority group. The strategy is efficient in improving the generalization and performance of models, especially when using imbalanced datasets.

e) Feature Selection: The feature selection facilitates the determination of the most relevant variables to be used in model training. Techniques such as as the ANOVA F-statistic, Chi-squared test and Mutual Information Feature Selection (MIFS) are employed to filter extraneous characteristics and therefore improving the effectiveness and accuracy of the model. Minimization of the number of features reduces the complexity of the model, which leads to a faster computation and better generalization.

iii) Training & Testing:

The data is separated into training and testing to determine performance of the model properly. The ratio used is 80:20 with 80 percent of the data being used in the model training and 20 percent on the model testing. This split ensures that this model has enough data to train on and at the same time a separate group of hidden data to validate itself. The split plays an important role in assessing the ability of the model to extrapolate and operate on new data that was not seen.

iv) Algorithms:

Naive Bayes [15] is used due to its simplicity and usefulness in working with large data sets. It applies the Bayes theorem to determine the risk of heart diseases based on feature probabilities, and it is highly effective with categorical data.

Support Vector Machine [20] (SVM) is used to determine the optimal hyperplane in separating different classes. It works remarkably better in high-dimensional space making it suitable to use in the complex combinations of features in prediction of heart disease.

XGBoost [17] is used because of its strong boosting power and enhances model correctness through iteration. It incorporates weak learners in a powerful predictive model thereby making it incredibly efficient in measuring heart disease risks.

AdaBoost [16] focuses on optimizing the imperfect classifiers, placing specific emphasis on the misclassified incidences. This is a predictive approach which improves predictative accuracy making it an important strategy in the classification of cardiac diseases in the model with certainty.

Bagging Classifier is used to reduce variance and enhance model stability [18]. When many models are integrated, forecasts that are made using different data subsets are combined to improve the accuracy of the estimates in heart disease risks.

Decision Tree The algorithm is applied due to its understandability and clearness. It divides the data based on the values of features and provides clear information about the decision-making to predict heart diseases [19].

K-Nearest Neighbors (KNN) is used because of its direct methodology in proximity based classification. It assesses the nearest data points to match the risk of heart diseases, based on similarity of occurrences.

Random Forest Combines a number of DT to enhance predictive accuracy and reduce overfitting. [14]. Such a group approach is effective in the prediction of cardiac disease and this has credible results in any body of data.

[16] Logistic Regression is used to model the probability of the appearance of heart diseases. It also measures the relationships between the dependent and the independent variables and this makes it suitable in binary classification tasks within the system.

[17] Voting Classifier combines the predictions of different models like Naive Bayes and SVM among others. This is an ensemble technique that enhances prediction accuracy because it uses the capabilities of various algorithms to classify heart disease.

Stacking Classifier combines predictions of a Boosted DT and ET with LightGBM. This hierarchical approach integrates multiple models, which make performance and accuracy more accurate in predicting the heart disease by detecting complex trends in the data.

4. RESULTS & DISCUSSION

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. The validity of a test refers to the ability of the test to make the right difference between patient and healthy cases. In order to determine the accuracy of a test, it will be necessary to compute the ratio of true positives and true negatives out of all the analyzed cases. This may be put mathematically as:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} (1)$$

Precision: Precision determines the rate of correctly identified cases/samples out of those that are detected as positive. Therefore, the precision calculation formula is defined in the following way:

$$Precision = \frac{\text{True Positive}}{\text{True Positive } + \text{False Positive}} (2)$$

Recall: In ML, recall is a measure to determine the ability of a model to identify all the relevant examples of a particular class. It is the proportion of the correctly predicted positive observations compared to the overall the real positive, which gives information on the effectiveness of a model in detecting events of a particular classification.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP}\ +\ \mathrm{FN}} (3)$$

F1-Score: The F1 score is used to measure the correctness of a ML model. It is a combination of precision and recall measures of a model. The measure of accuracy measures how often a model makes correct predictions when applied on the whole dataset.

$$F1\,Score = 2 * \frac{Recall \text{ X } Precision}{Recall + Precision} * 100 (4)$$

Tables (1, 2 & 3) evaluate the performance indicators, namely accuracy and precision, recall, and F1-score of each algorithm. The Stacking

Classifier wins all metrics by a significant margin over every other algorithm. The tables present a comparative analysis of the measures of the alternative methods.

Table.1 Performance Evaluation Metrics for Anova FS

Model	Accuracy	Precision	Recall	F1 Score
Naive Bayes	0.848	0.850	0.848	0.849
SVM	0.682	0.686	0.682	0.682
XGBoost	0.818	0.820	0.818	0.818
AdaBoost	0.833	0.845	0.833	0.834
Bagging	0.818	0.826	0.818	0.819
Decision Tree	0.788	0.790	0.788	0.788
KNN	0.727	0.729	0.727	0.728
Random Forest	0.864	0.868	0.864	0.864
Logistic Regression	0.864	0.864	0.864	0.864
Voting	0.818	0.818	0.818	0.818
Stacking	**1.000**	**1.000**	**1.000**	**1.000**

Graph.1 Comparison Graphs for Anova FS

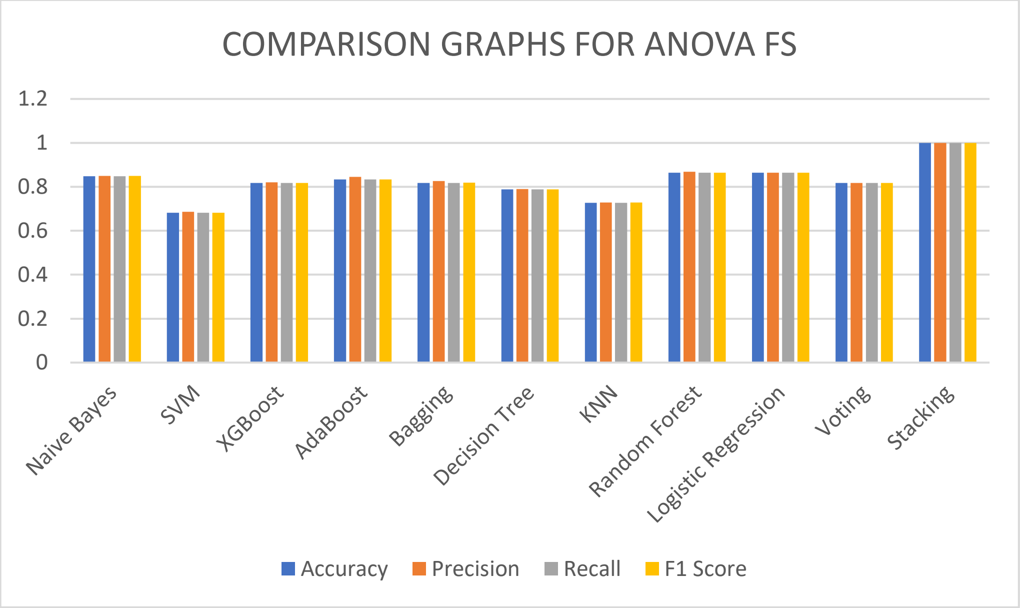


Table.2 Performance Evaluation Metrics for Chi2 FS

Model	Accuracy	Precision	Recall	F1 Score
Naive Bayes	0.788	0.790	0.788	0.788
SVM	0.652	0.656	0.652	0.652
XGBoost	0.818	0.835	0.818	0.820
AdaBoost	0.773	0.773	0.773	0.773
Bagging	0.803	0.815	0.803	0.804
Decision Tree	0.727	0.735	0.727	0.728
KNN	0.621	0.622	0.621	0.621
Random Forest	0.879	0.886	0.879	0.879
Logistic Regression	0.803	0.807	0.803	0.803
Voting	0.818	0.818	0.818	0.818

Stacking	**1.000**	**1.000**	**1.000**	**1.000**

Graph.2 Comparison Graphs for HHO FS in Chi2 FS

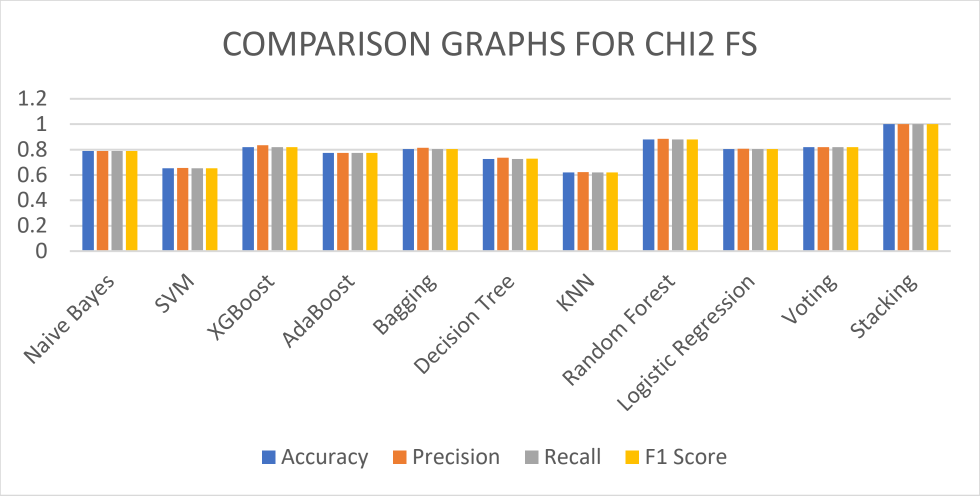
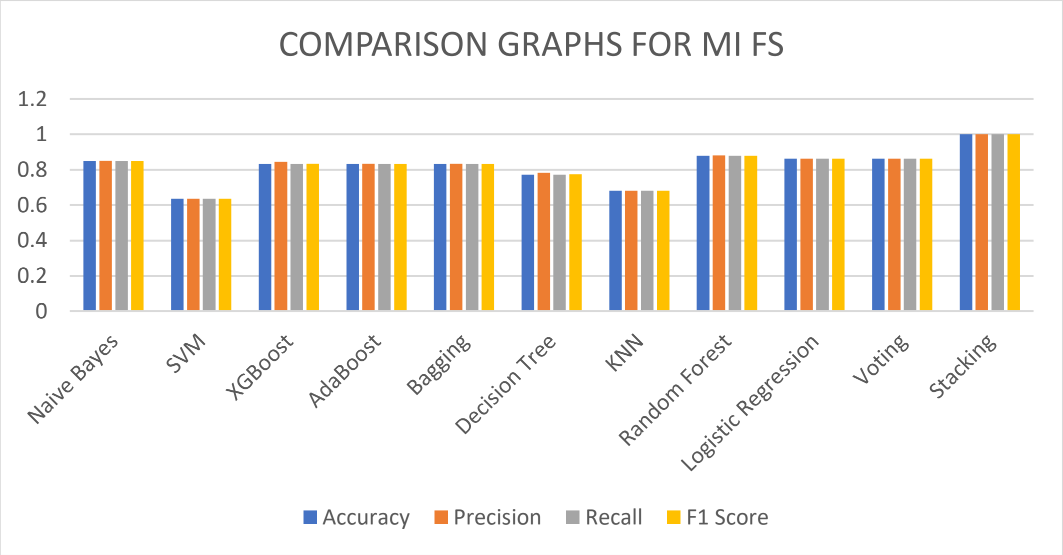


Table.3 Performance Evaluation Metrics for MI FS

Model	Accuracy	Precision	Recall	F1 Score
Naive Bayes	0.848	0.850	0.848	0.849
SVM	0.636	0.636	0.636	0.636
XGBoost	0.833	0.845	0.833	0.834
AdaBoost	0.833	0.834	0.833	0.833
Bagging	0.833	0.834	0.833	0.833
Decision Tree	0.773	0.784	0.773	0.774
KNN	0.682	0.682	0.682	0.682
Random Forest	0.879	0.881	0.879	0.879
Logistic Regression	0.864	0.864	0.864	0.864
Voting	0.864	0.864	0.864	0.864
Stacking	**1.000**	**1.000**	**1.000**	**1.000**

Graph.3 Comparison Graphs for MI FS



The depictions of accuracy, precision, recall and F1-Score are light blue, orange, grey, and light yellow respectively as in ***Graphs 1, 2, and 3***. The Stacking Classifier, in comparison to the rest of the models, has shown to perform better in all measures with the highest values. These findings have been presented in the above graphs.

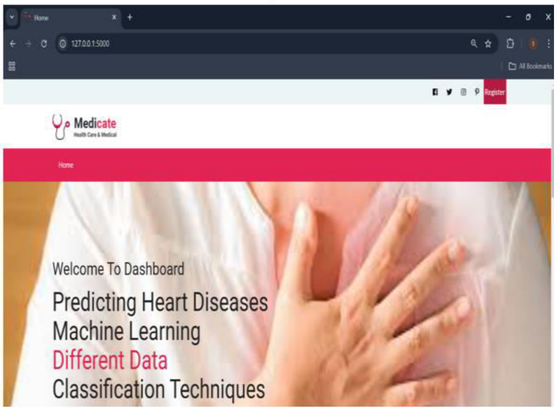


Fig.3 Home Page

Figure 3 represents a user interface dashboard with an explanation that allows navigating the site.

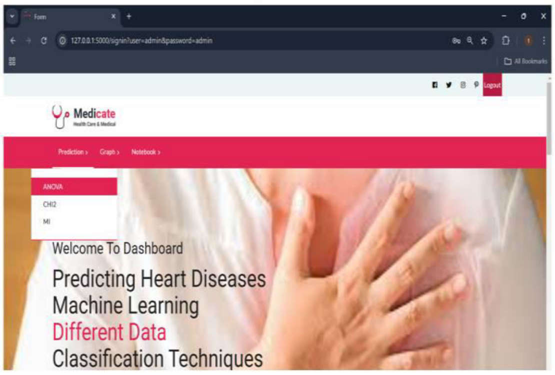
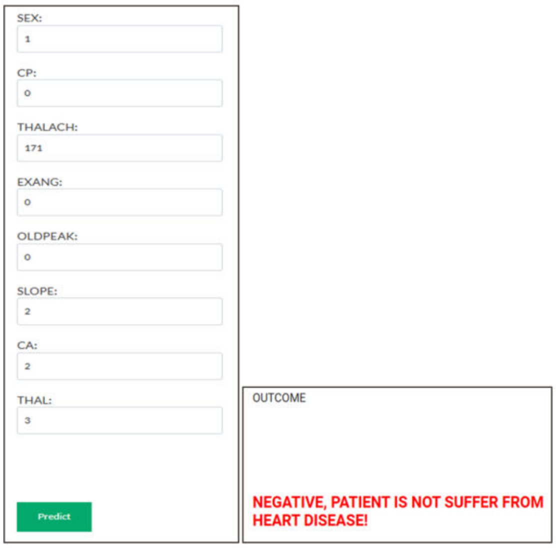


Fig.4 ANOVA dataset loading

Figure 4 shows a user input page, which allows one to submit an ANOVA dataset to test it.



SEX: 1

CP: 0

THALACH: 171

EXANG: 0

OLDPEAK: 0

SLOPE: 2

CA: 2

THAL: 3

OUTCOME

NEGATIVE, PATIENT IS NOT SUFFER FROM HEART DISEASE!

Fig.5 Test result

A results screen is shown in figure 5 where the user is provided with the output of the input data that has been loaded.

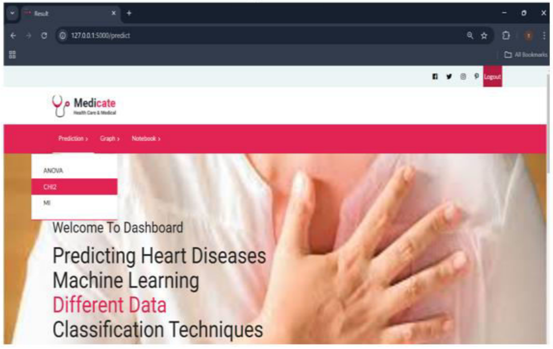
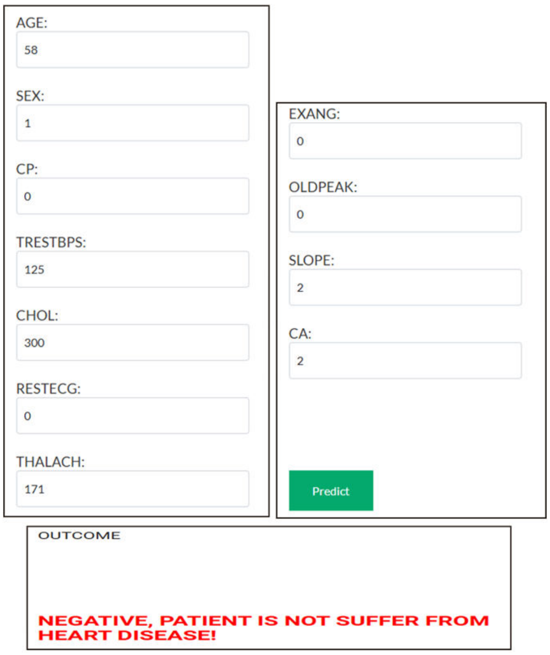


Fig.6 CHI2 dataset loading

Figure 6 depicts a user input interface, which allows the user to upload the CHI2 dataset to do a test.



AGE: 58

SEX: 1

CP: 0

TRESTBPS: 125

CHOL: 300

RESTECG: 0

THALACH: 171

EXANG: 0

OLDPEAK: 0

SLOPE: 2

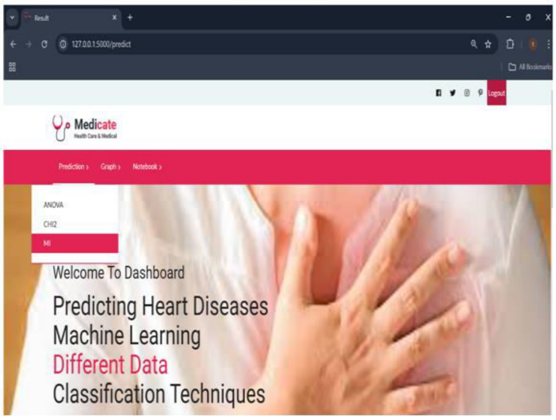
CA: 2

OUTCOME

NEGATIVE, PATIENT IS NOT SUFFER FROM HEART DISEASE!

Fig.7 Test result

The results screen as shown in figure 7 gives the user an output of the input data that has been loaded.

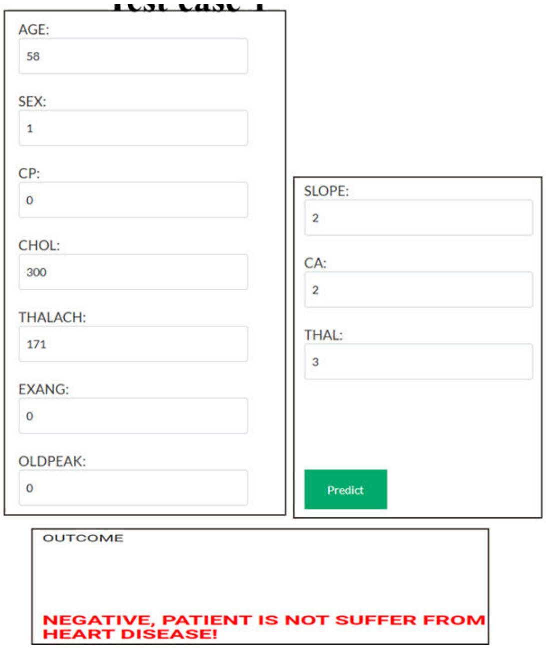


Welcome To Dashboard

Predicting Heart Diseases Machine Learning Different Data Classification Techniques

Fig.8 MI dataset loading

Figure 8 shows a user input interface, as it allows uploading the MI dataset to run the test.



AGE: 58

SEX: 1

CP: 0

CHOL: 300

THALACH: 171

EXANG: 0

OLDPEAK: 0

SLOPE: 2

CA: 2

THAL: 3

OUTCOME

NEGATIVE, PATIENT IS NOT SUFFER FROM HEART DISEASE!

Fig.9 Test result

Figure 9 shows a result screen in which the user is provided with the data of the input data uploaded.

5. CONCLUSION

To sum up, the proposed solution demonstrates the effectiveness of the use of high-level ML strategies in cardiac disease prediction. By applying feature selection methods like the ANOVA F-statistic, Chi-squared test and Mutual Information, the system identifies the most important predictors efficiently therefore enhancing the overall performance of the model. The application of SMOTE in correcting the imbalance of classes increases the ability of the model to detect cases of heart diseases, therefore making the predictions equitable and reliable.

Among the algorithms that have been tested, the Stacking Classifier, which combines Boosted DT, ET, and LightGBM, achieved the best performance with an outstanding accuracy of the full set of feature selection methods of 100 percent. This result shows the effectiveness of the ensemble techniques in combining the benefits of individual classifiers to promote the precision of the forecasts. The proposed approach improves the correct and timely diagnosis of cardiovascular conditions by combining the latest feature selection and the complex ensemble learning, proving its usefulness in the clinical practice and medical decision-making.

Third, the future study can explore the possibility of additional methodologies, e.g. deep learning models and neural networks, to improve predictive accuracy. More improvements could be given by using a more advanced ensemble method, like Gradient Boosting, or stacking on a wider range of base classifiers. An addition of supplementary feature selection tools, e.g. Recursive Feature Elimination (RFE) or L1 regularization, may increase the effectiveness of the model. Exploring the time-series data and incorporating the time variable could provide deeper information to predict the outcomes of heart diseases.

REFERENCES

[1] (2023). World Health Organization. Cardiovascular Diseases (CVDs). Accessed: May 5, 2023. [Online]. Available: https://www.afro.who.int/ health-topics/cardiovascular-diseases

[2] Z. Alom, M. A. Azim, Z. Aung, M. Khushi, J. Car, and M. A. Moni, ‘‘Early stage detection of heart failure using machine learning techniques,’’ in Proc. Int. Conf. Big Data, IoT, Mach. Learn., Cox’s Bazar, Bangladesh, 2021, pp. 23–25.

[3] S. Gour, P. Panwar, D. Dwivedi, and C. A. Mali, ‘‘Machine learning approach for heart attack prediction,’’ in Intelligent Sustainable Systems. Singapore: Springer, 2022, pp. 741–747.

[4] C. Gupta, A. Saha, N. S. Reddy, and U. D. Acharya, ‘‘Cardiac disease prediction using supervised machine learning techniques,’’ J. Phys., Conf. Ser., vol. 2161, no. 1, 2022, Art. no. 012013.

[5] K. Shameer, ‘‘Machine learning predictions of cardiovascular disease risk in a multi-ethnic population using electronic health record data,’’ Int. J. Med. Inform., vol. 146, Feb. 2021, Art. no. 104335.

[6] M. Yang, X. Wang, F. Li, and J. Wu, ‘‘A machine learning approach to identify risk factors for coronary heart disease: A big data analysis,’’ Comput. Methods Programs Biomed., vol. 127, pp. 262–270, Apr. 2016.

[7] C. Ngufor, A. Hossain, S. Ali, and A. Alqudah, ‘‘Machine learning algorithms for heart disease prediction: A survey,’’ Int. J. Comput. Sci. Inf. Secur., vol. 14, no. 2, pp. 7–29, 2016.

[8] A. Farag, A. Farag, and A. Sallam, ‘‘Improving heart disease prediction using boosting and bagging techniques,’’ in Proc. Int. Conf. Innov. Trends Comput. Eng. (ITCE), Mar. 2016, pp. 90–96.

[9] X. Zhang, Y. Zhang, X. Du, and B. Li, ‘‘Application of XGBoost algorithm in clinical prediction of coronary heart disease,’’ Chin. J. Med. Instrum., vol. 43, no. 1, pp. 12–15, 2019.

[10] Y. Liu, X. Li, and J. Ren, ‘‘A comparative analysis of machine learning algorithms for heart disease prediction,’’ Comput. Methods Programs Biomed., vol. 200, Nov. 2021, Art. no. 105965.

[11] N. S. Hussein, A. Mustapha, and Z. A. Othman, ‘‘Comparative study of machine learning techniques for heart disease diagnosis,’’ Comput. Sci. Inf. Syst., vol. 17, no. 4, pp. 773–785, 2020.

[12] S. Akbar, R. Tariq, and A. Basharat, ''Heart disease prediction using different machine learning approaches: A critical review,'' J. Ambient Intell. Humanized Comput., vol. 11, no. 5, pp. 1973–1984, 2020.

[13] A. Zarshenas, M. Ghanbarzadeh, and A. Khosravi, ''A comparative study of machine learning algorithms for predicting heart disease,'' Artif. Intell. Med., vol. 98, pp. 44–54, Oct. 2019.

[14] I. Kaur G. Singh, ''Comparative analysis of machine learning algorithms for heart disease prediction,'' J. Biomed. Inform., vol. 95, Jul. 2019, Art. no. 103208.

[15] Y. Li, W. Jia, and J. Li, ''Comparing different machine learning methods for predicting heart disease: A telemedicine case study,'' Health Inf. Sci. Syst., vol. 6, p. 7, Dec. 2018.

[16] X. Zhang, Y. Zhou, and D. Xie, ''Heart disease diagnosis using machine learning and expert system techniques: A survey paper,'' J. Med. Syst., vol. 42, no. 7, p. 129, 2018.

[17] J. Wu, J. Roy, and W. F. Stewart, ''A comparative study of machine learning methods for the prediction of heart disease,'' J. Healthcare Eng., vol. 2017, Jan. 2017, Art. no. 7947461.

[18] Z. Ahmed, K. Mohamed, and S. Zeeshan, ''Comparison of machine learning algorithms for predicting the risk of heart disease: A systematic review,'' J. Healthcare Eng., vol. 2016, Jan. 2016, Art. no. 7058278.

[19] X. Chen, Z. Hu, and Y. Cao, ''Heart disease diagnosis using decision tree and naïve Bayes classifiers,'' World Congr. Medical Phys. Biomed. Eng., vol. 14, pp. 1668–1671, Aug. 2007.

[20] N. Samadiani, A. M. E Moghadam, and C. Motamed, ''SVM-based classification of cardiovascular diseases using feature selection: A high-dimensional dataset perspective,'' J. Med. Syst., vol. 40, no. 11, p. 244, Nov. 2016.