Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 4, No.1, April : 2025

CARDIOCARE: ML-BASED HEART DISEASE PREDICTION

J.Vidhyajanani, Assistant Professor, Department of Computer Science and Engineering, Paavai College of Engineering, Namakkal.
R.Bharathkumar[a],G.Vigneshwaran[b], S.Lingesh raji[c], M.Pradeep Raj[d], K.Kesavan[e], R.Sriman Abinesh[f], V.Pradeep[g] , Students, Department of CSE(Artificial Intelligence and Machine Learning), Paavai College of Engineering, Namakkal

ABSTRACT:
Over the past several decades, heart-related illnesses, often known as cardiovascular diseases (CVDs), have been the leading cause of mortality worldwide and are now considered to be the most deadly condition, not just in India but globally. Accurately estimating a patient's risk of heart disease is a crucial problem in medical data analysis as it is essential for early intervention and lowering death rates. Early detection enables prompt treatment and ongoing healthcare provider monitoring, both of which are crucial but frequently constrained by the incapacity of medical personnel to continuously supervise patients. Physicians can lower fatality rates by detecting cardiac issues early and keeping a close eye on their patients. Heart disease detection is not always reliable, and doctors cannot be in regular contact with their patients. To find the machine learning system that produces the best accurate forecasts of heart illness, we employed the Synthetic Minority Oversampling Technique (SMOTE) to remove inconsistent data. The suggested approach might help medical professionals swiftly and affordably detect early-stage cardiac disease. Based on the input symptoms, we have developed a smartphone application that uses the best machine learning algorithm to forecast heart disease in real time. The experimental findings showed that the SF-2 feature subset and the combined datasets yielded the best results from the XGBoost algorithm. It had a 92.68% F1 score, a 98% AUC, 97.57% accuracy, 96.61% sensitivity, 90.48% specificity, and 95.00% precision. In order to comprehend how the system arrives at its final predictions, we have created an explainable AI technique based on SHAP techniques.
Keywords: Machine Learning, Heart Disease, Ml algorithm, SMOTE.

INTRODUCTION:

Heart conditions are the biggest cause of mortality worldwide [1]. According to a World Health Organization estimate, heart disease and stroke cause 17.5 million deaths globally each year. The majority of the more than 75% of heart disease-related fatalities take place in middle- and low-income nations. Furthermore, 80% of all deaths from CVDs are attributable to heart attacks and strokes [2]. Heart disease is frequently diagnosed after a physical examination and observation of the patient's symptoms. Among the risk factors for cardiovascular disease include smoking, advanced age, a family history of heart disease, high blood pressure, obesity, diabetes, stress, high cholesterol, and inactivity [3]. Some of these risk factors may be decreased by changing one's lifestyle to include things like exercising, reducing stress, stopping smoking, and decreasing weight. Medical history, physical examination, and imaging tests such cardiac MRIs, echocardiograms, electrocardiograms, and blood tests are used to identify heart disease. Heart disease can be treated with medications, lifestyle changes, coronary artery bypass surgery, angioplasty, or implanted devices like pacemakers or defibrillators [4].

People's lifestyle choices, inactivity, and intake of processed foods are the primary causes of the notable rise in heart disease in recent years. Advanced heart illness can result in heart attacks and put patients' lives in jeopardy, therefore it's critical to identify the condition early and promptly using therapeutic and intelligent techniques. Patients' unwillingness to take part in clinical trials is one of the main obstacles to diagnosing cardiac disease. However, these studies are expensive and time-consuming, which is why they are not given much attention. Certain techniques may be used to

examine the pattern of cardiac disease by examining data from patients and healthy individuals, in contrast to clinical approaches for diagnosing the condition [5]. Information on patients with medical reports is currently easily accessible in databases and is expanding daily in the healthcare industry. This uneven raw data is extremely redundant. Pre-processing is necessary to increase classification performance, shorten training algorithm execution times, and extract significant features [6].
Recent developments in computing power and machine learning's reprogramming capabilities enhance these procedures and create new avenues for healthcare research, particularly in the area of early disease prediction to increase survival rates, such as in cancer and cardiovascular disease. Applications for machine learning are numerous, ranging from determining disease risk factors to creating cutting-edge automotive safety systems. The study's goal is to offer an ML method for predicting heart disease. We tested machine learning methods on huge, publicly available datasets for heart disease prediction. Building a novel machine learning method that can accurately identify several high-definition datasets is the goal of this project. Its performance will then be compared to that of existing excellent models.
This study's usage of a private HD dataset is one of its main contributions. Between 2022 and 2025, 200 data samples were freely submitted by Egyptian specialty hospitals. From these individuals, we were able to gather about 11 attributes. The urgent need for early HD prediction in Saudi Arabia and Egypt, where the HD rate is rising quickly, is the focus of this effort. The performance of the suggested model was assessed by the authors using ML classification algorithms on a combined dataset that included both private and CHDD datasets. This method predicts HD correctly using a merged dataset. In contrast to previous research, this approach is novel. The declared objective of the study was to use the SF-2 feature subset and the pooled datasets to predict HD. Hyperparameters were also used to optimize the machine learning techniques used in this paper. For every ML classifier, we have adjusted the hyperparameters. Using the combined datasets and the SF-2 feature subset, the suggested technique achieved 95.57% accuracy rates with optimal hyperparameters.

LITERATURE REVIEW:
Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), SVM with grid search (SVMG), K-Nearest Neighbor (KNN), and Naïve Bayes (NB) are among the machine learning models that Dubey A. K. et al. studied for heart disease detection. Training and testing were conducted using the UCI Machine Learning repository's Cleveland and Statlog datasets. LR and SVM classifier models outperform the Cleveland dataset with 89% accuracy, according to the experimental results, while LR outperforms the Statlog dataset with 93% accuracy [7]. Karthick K. et al. developed an ML model for predicting the risk of heart disease using the SVM, Gaussian Naive Bayes (GNB), LR, LightGBM, XGBoost, and RF algorithms. To choose the top characteristics from the Cleveland heart disease dataset, the authors of this study used the Chi-square statistical test. The RF classifier model had the greatest classification accuracy rate of 88.5% following feature selection [8].
Using the Cleveland heart disease dataset, Veisi H. et al. created a number of machine learning (ML) models, including DT, RF, SVM, XGBoost, and Multilayer Perceptron (MLP), to predict heart disease. The dataset was subjected to a number of feature selection and preparation procedures, including outlier identification and normalization. The MLP has the best accuracy of 94.6% of the ML models that were assessed [9]. Sarra R. R. et al. used the Cleveland and Statlog datasets from the UCI Machine Learning repository to propose a novel classification model based on SVM for improved heart disease prediction. To increase the model's prediction accuracy, the $\chi 2$ statistical optimum feature selection approach was applied. Using a variety of performance indicators, the suggested model's performance is compared to that of conventional classifier models. The findings indicate that the accuracy increased from 85.29% to 89.7% when the suggested model was used [10].

Using 297 records and 13 characteristics from the Cleveland dataset, S. Mohan et al. created an efficient hybrid random forest with a linear model (HRFLM) to improve the prediction accuracy of heart disease. They found that the best error rates were obtained using the LM and RF approaches. Using Orange and Weka data mining tools, S. Kodati et al. created a heart disease prediction system (HDPS) with the Cleveland dataset, which consisted of 297 instances and 13 characteristics. They then assessed the accuracy and recall metrics for the naïve Bayes, SMO, RF, and KNN classifiers.

It is clear from the experimental studies that feature selection and data pre-processing may significantly improve machine learning algorithms' classification accuracy. To ensure that the dataset was complete, the majority of researchers used the mean value or the majority mark of that characteristic to replace the missing values during pre-processing. The missing valued occurrences were eliminated in some works. The vast exploring space makes feature selection a difficult chore. Depending on how many characteristics are present in the dataset, it increases exponentially. An efficient, all-encompassing search strategy is needed for feature selection in order to address this problem. In order to improve prediction accuracy, several research have also used ensemble models, which mix many fundamental learning techniques.

DATASET AND FEATURE SELECTION:

In order to forecast heart disease, this study uses both the CHDD and a private dataset. There are 200 examples in the private dataset compared to 303 in the CHDD dataset, and both share the same characteristics. There are 13 attributes (demographic, clinical, and laboratory factors) linked to each of the 503 entries in the merged dataset. Numerous characteristics of the datasets, such as age, gender, blood pressure, cholesterol levels, electrocardiogram readings (ECG), chest pain, exercise-induced angina, blood sugar levels during fasting, maximum heart rate attained, oldpeak, coronary artery, thalassemia, and other clinical and laboratory measurements, can be used to predict heart disease.

Preprocessing was done on the data that was gathered for this study. There are two incorrect TS entries and four erroneous CMV data in the CHDD. To reflect the optimal values for every field, inaccurate data is rectified. After that, StandardScaler is used to normalize every feature to the appropriate coefficient, guaranteeing that every feature has a single variance and zero mean. An ordered and constructed enhanced dataset was selected by taking into account the patient's history of cardiac issues as well as other medical concerns.

This study's dataset consists of a combination of publicly available WBCD and selected private datasets. We may apply the holdout validation approach by dividing the two datasets in this manner. In this study, the test dataset has 25% of the data, whereas the training dataset contains 75%. This study measures the interdependence of variables using the mutual information approach. Greater dependence and information collecting are shown by larger numbers. The significance of features offers important information about each feature's applicability and predictive ability within a dataset. As shown in Fig. 1, the thalach feature receives the highest value of 13.65% using this reciprocal information approach, while the fbs feature receives the lowest importance of 1.91%.

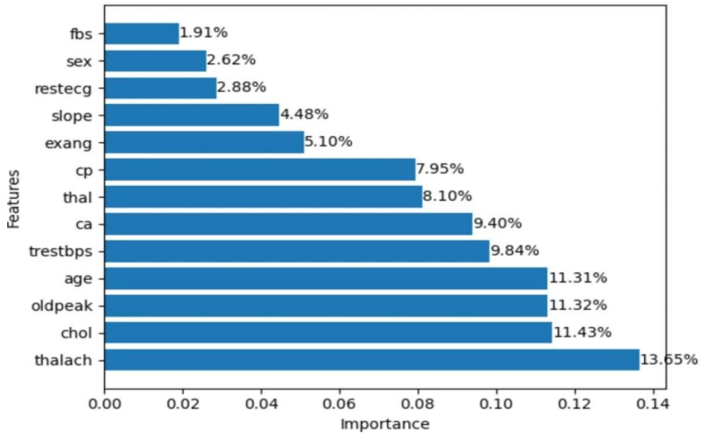

Fig 1: Importance features of Dataset

In this study, we use the Python Scikit-learn package to pick and classify features []. A variety of machine learning classifiers, including RF, LR, KNN, bagging, DT, AdaBoost, XGBoost, SVM, voting, and Naive Bayes, were first used to examine the processed dataset. These classifiers were then assessed for overall accuracy. The second phase involved creating heat maps of correlation matrices and other visualizations of correlations between various data sets using the Seaborn modules in Python. Thirdly, a broad range of feature selection techniques (FSM) were used, including mutual information (MI), chi-square, and analysis of variance (ANOVA). We selected the XGBoost classifier with SMOTE utilizing the combined datasets and SF-2 feature subset based on the research's evaluation of performance criteria (see Table 6). We will use a range of integrated development environments (IDEs), such as Android Studio 14.0, Python 3.10, Spyder, Java 11, and Pickle 5, to deploy the model and incorporate the most accurate method in a mobile application.

SMOTE ALGORITHM :

The dataset is frequently unbalanced in various prediction tasks, including medical diagnosis and prediction. This suggests that one class is underrepresented compared to the other, usually the class of interest. SMOTE generates synthetic minority class samples by interpolating minority class examples. The prediction model is given enough minority class instances to learn from thanks to this balanced class distribution. Predictive models that employ an unbalanced dataset may be biased in favor of the majority class, which might result in subpar predictions for the minority class. A major problem is accurately predicting the minority class. Using SMOTE improves accuracy and predictive performance, especially for the minority class, by training the model on a more balanced dataset. In applications where excluding the minority class (disease cases, for example) might have serious repercussions, this is crucial.

When trained on unbalanced datasets, predictive models frequently show low recall for the minority class but good precision for the majority class. This implies that even when they do identify minority class occurrences, they overlook a significant percentage of them. SMOTE creates a more balanced and efficient model by increasing recall without compromising precision. Practically speaking, this indicates that the model is more adept at locating all pertinent examples rather than just a few.

Biased models can produce unjust results in prediction applications, particularly when the minority class is underrepresented. SMOTE reduces this bias by training the model with a sufficient number of minority class samples. This contributes to the development of a more egalitarian model that predicts outcomes more fairly for all groups. When training on unbalanced data, models may perform well on the majority class, but they struggle to generalize to new, unknown data, especially for the minority class. The model is better able to generalize its predictions to new data when a

balanced training set is created using SMOTE, which results in more dependable and consistent performance in real-world applications.
Robustness is crucial for deployed machine learning programs. Predictive models frequently encounter unbalanced or distorted real-world data. SMOTE lowers the chance of failure in production settings by assisting in the development of a more resilient model that can manage such data more skillfully. For applications like predictive maintenance, where detecting infrequent but significant failures can avoid expensive downtime, this is essential.

EXPERIMENTAL RESULT:
To forecast cardiac disorders from a dataset, we utilize Jupyter Notebook 7. It makes document production, including live coding, easier and makes it easier to visualize the many data relationship graphs in the dataset. Cleaning the CHDD with Python's Pandas and NumPy modules (version 24.2.0) is the first stage in this study. The dataset34 is then preprocessed using Python's Scikit-learn module's StandardScaler function. Three sets of features (SF) are produced in the second stage of the procedure, which uses a feature selection technique to determine each feature's relevance. Thirdly, we divided the dataset into sets for testing and training. Of the data, we utilize 25% for testing and 75% for training.
When it came to accuracy, A4's SF-2 accuracy calculation was the most exact (97.57%), followed by its SF-1 and SF-3 accuracy estimates (93.17% and 94.19%), respectively. A9 came in second place with an accuracy of 93.07% across all three SFs. However, A5 found that out of all the classifiers, SF-1 and SF-3 had a poor accuracy of 85.15%. The accuracy of A3 and A10 for SF-2 and SF-3 was also poor, at 86.14% and 86.12%, respectively. The accuracy of the other approaches ranges from 87.13% to 90.00%. Additionally, this result indicates that the SF-2-based XGBoost algorithm approach is the most efficient way to process the dataset.

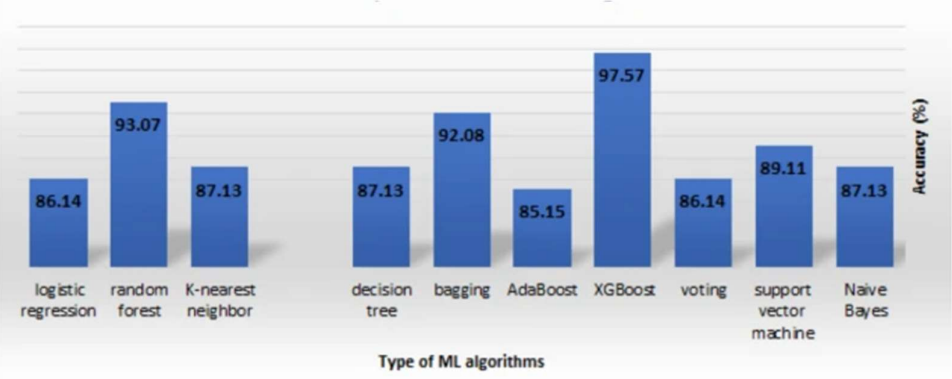

Fig2 provides a summary of the specificity analysis we conducted on each of these methods. The lowest score for SF-2 and FS-3 was 73.81% for A3. A4 and A9 had the highest scores (90.48%) of any SFs, according to the analysis's findings. The best score with SF-3 alone was obtained with an A7 for SF-3 (92.86%), in comparison to the outcomes of the other procedures.
The aim of this work is to predict HD using ML classifiers. The results of the trial demonstrated that the XGBoost algorithm had the highest prediction accuracy for HD incidence. The mutual information-based feature selection technique classifies the following characteristics as crucial for HD prediction: thalach, chol, oldpeak, age, trestbps, ca, thal, cp, exang, slope, and fbs. Using the gathered data, we have optimized the hyperparameters and oversample using the SMOTE approach. The best results were obtained using the XGBoost approach with SMOTE. Using the pooled datasets, the study was able to predict HD: 97.57% accuracy, 96.61% sensitivity, 90.48% specificity, 95.00% precision, 92.68% F1 score, and 98% AUC were the experimental outcomes.

CONCLUSION:

Ten different machine learning approaches using SMOTE were used to the characteristics that were chosen using a variety of methods in this study. We were able to determine the most important characteristics that are very good in predicting heart disease thanks to this procedure. Each algorithm uses a different mix of characteristics to provide a distinct score. ANOVA, chi-square, and MI were the three techniques we employed to choose characteristics. Three chosen feature groups—SF-1, SF-2, and SF-3—were subjected to these techniques. The best model and feature subset were identified using ten machine learning classifiers. Naive Bayes, SVM, voting, XGBoost, AdaBoost, bagging, DT, KNN, RF, and LR were the classifiers that were employed. To assess the chosen algorithms and gauge the performance accuracy of the heart disease detection system, we used a popular open-access dataset and many cross-validation procedures. The performance of XGBoost was more significant than that of any other algorithms. With 97.64% accuracy, 96.61% sensitivity, 90.48% specificity, 95.00% precision, a 92.68% F1 score, and a 98% AUC, the XGBoost classifier outperformed the others when applied to the SF-2 feature subset. Using SHAP approaches, we created an explainable AI approach to comprehend the system's result prediction process.

Last but not least, the authors are creating a smartphone app that will enable users to input symptoms and make fast and precise predictions about cardiac disease. In order to anticipate heart illness and provide the detection result immediately, we will include the best XGBoost technology into the mobile app. Since the mobile app predicts cardiac disease based on symptoms, we will take into account and deal with the influence of "dark data" throughout its deployment. Dark data is information that is available but is either underused or not collected because of poor reporting, ignorance, or constraints in data collecting.

REFERENCES:

1. World Health Organization. Cardiovascular Diseases (CVDs). Available online: (2023
2. Alom, Z. et al. Early Stage Detection of Heart Failure Using Machine Learning Techniques. In Proceedings of the International Conference on Big Data, IoT, and Machine Learning, Cox's Bazar, Bangladesh, 23–25 September (2021).
3. Gour, S., Panwar, P., Dwivedi, D. & Mali, C. A machine learning approach for heart attack prediction. In Intelligent Sustainable Systems (eds Nagar, A. K., Jat, D. S., Marín-Raventós, G. & Mishra, D. K.) 741–747 (Springer, Singapore, 2022). Nguyen T., Wang Z.A. Cardiovascular screening and early detection of heart disease in adults with chronic kidney disease. J. Nurse Pract. 2019;15:34–40. doi: 10.1016/j.nurpra.2018.08.004.
4. Pasha, S.J.; Mohamed, E.S. Novel Feature Reduction (NFR) Model with Machine Learning and Data Mining Algorithms for Effective Disease Risk Prediction. IEEE Access 2020, 8, 184087–184108.
5. Swain, D.; Pani, S.K.; Swain, D. A Metaphoric Investigation on Prediction of Heart Disease using Machine Learning. In Proceedings of the 2018 International Conference on Advanced Computation and Telecommunication, ICACAT, Bhopal, India, 28–29 December 2018; pp. 1–6.
6. Dubey A.K., Choudhary K., Sharma R. Predicting Heart Disease Based on Influential Features with Machine Learning. Intell. Autom. Soft Comput. 2021;30:929–943. doi: 10.32604/iasc.2021.018382.
7. Karthick K., Aruna S.K., Samikannu R., Kuppusamy R., Teekaraman Y., Thelkar A.R. Implementation of a heart disease risk prediction model using machine learning. Comput. Math. Methods Med. 2022;2022:6517716. doi: 10.1155/2022/6517716.
8. Veisi H., Ghaedsharaf H.R., Ebrahimi M. Improving the Performance of Machine Learning Algorithms for Heart Disease Diagnosis by Optimizing Data and Features. Soft Comput. J. 2021;8:70–85.

9. Sarra R.R., Dinar A.M., Mohammed M.A., Abdulkareem K.H. Enhanced heart disease prediction based on machine learning and χ2 statistical optimal feature selection model. Designs. 2022;6:87. doi: 10.3390/designs6050087.
10. Singh A., Kumar R. Heart disease prediction using machine learning algorithms; Proceedings of the 2020 International Conference on Electrical and Electronics Engineering (ICE3); Gorakhpur, India. 14–15 February 2020; pp. 452–457.
11. Sahoo G.K., Kanike K., Das S.K., Singh P. Machine Learning-Based Heart Disease Prediction: A Study for Home Personalized Care; Proceedings of the 2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP); Xi'an, China. 22–25 August 2022; pp. 1–6.
12. Khdair H. Exploring Machine Learning Techniques for Coronary Heart Disease Prediction. 2021. [(accessed on 12 April 2023)].
13. Ahmad G.N., Fatima H., Abbas M., Rahman O., Alqahtani M.S. Mixed machine learning approach for efficient prediction of human heart disease by identifying the numerical and categorical features. Appl. Sci. 2022;12:7449. doi: 10.3390/app12157449.
14. Chou J.-S., Truong D.-N. A novel metaheuristic optimizer inspired by behavior of jellyfish in ocean. Appl. Math. Comput. 2021;389:125535. doi: 10.1016/j.amc.2020.125535
15. Acharya U.R., Fujita H., Oh S.L., Raghavendra U., Tan J.H., Adam M., Gertych A., Hagiwara Y. Automated identification of shockable and non-shockable life-threatening ventricular arrhythmias using convolutional neural network. Futur. Gener. Comput. Syst. 2018;79:952–959. doi: 10.1016/j.future.2017.08.039.
16. 27.Yao Q., Wang R., Fan X., Liu J., Li Y. Multi-class arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network. Inf. Fusion. 2020;53:174–182. doi: 10.1016/j.inffus.2019.06.024.