Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques

Prof. Rakesh Ramesh Tannu
H. O. D. Information Technology, JCEI's Jaihind Polytechnic Kuran,

***Abstract*: Heart disease (HD), including heart attacks, is a primary cause of death across the world. In the area of medical data analysis, one of the most difficult problems to solve is determining the probability of a patient having heart disease. Mortality rates can be reduced through early detection of heart diseases and continuous monitoring of patients by doctors Regretfully, a doctor cannot always be in contact with a patient, and heart illness cannot always be appropriately diagnosed. By offering a more accurate foundation for forecasting and decision-making based on data provided by healthcare sectors worldwide, machine learning (ML) holds promise for assisting in diagnosis. In order to create a precise machine learning strategy for predicting heart disease in its early stages, this study will use a number of feature selection techniques. Three different techniques—chi-square, analysis of variance (ANOVA), and mutual information (MI)—were used in the feature selection procedure. The three feature groups that were ultimately selected were referred to as SF-1, SF-2, and SF-3, respectively. Then, ten different ML classifiers were used to determine the best technique, and which feature subset was the greatest fit. These classifiers included Naive Bayes, support vector machine (SVM), voting, XGBoost, AdaBoost, bagging, decision tree (DT), K-nearest neighbor (KNN), random forest (RF), and logistic regression (LR), and they were denoted as (A1, A2, . . . , A10). The proposed approach for predicting heart diseases was evaluated using a private dataset, a publicly available dataset, and multiple cross-validation methods To find the classifier that generates the best rate of accurate heart disease predictions, we applied the Synthetic Minority Oversampling Technique (SMOTE) to fix the issue of unbalanced data. To comprehend how the system forecasts its final outcomes, an explainable artificial intelligence strategy utilizing SHAP techniques is being developed. With its low cost and short turnaround time, the suggested method showed tremendous potential for the healthcare industry in predicting early-stage cardiac disease. In the end, the most effective machine learning technique was applied to create a mobile application that enables users to input HD symptoms and promptly obtain a heart disease forecast. The proposed technique had great promise for the healthcare sector to predict early-stage heart disease with cheap cost and minimal time. Ultimately, the best ML method has been used to make a mobile app that lets users enter HD symptoms and quickly receive a heart disease prediction.**

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

I. INTRODUCTION

The heart is a muscular organ that serves as the circulatory system's main pumping organ. It is a part of the cardiovascular system and is in charge of pumping blood throughout the body. The system of arteries, veins, and capillaries that transport blood throughout the body is also referred to as the "cardiovascular system." Disturbances in the regular outflow of blood from the circulatory system cause a number of different types of heart disease that collectively are referred to as cardiovascular disease (CVD) are brought on by disruptions in the regular outflow of blood from the circulatory system.

On a global scale, heart diseases are continuously rated as the primary reason for people's death [1]. Heart disease and stroke account for 17.5 million annual deaths worldwide, according to the World Health Organization's report. More than 75% of deaths caused by heart diseases take place mostly in nations with middle and low income. In addition, heart attacks and strokes are responsible for 80 percent of all fatalities caused by CVDs [2]. Every person should be happy and healthy, according to Sustainable Development Goal (SDG) 3 of the UN. This study looks into cardiovascular illness. A physical examination and observation of the patient's symptoms are frequently used to identify heart disease. Among the risk factors for cardiovascular disease are high blood pressure, obesity, diabetes, stress, smoking, age, family history of heart disease, high cholesterol, and insufficient physical activity. A physical examination and

observation of the patient's symptoms are frequently used to identify heart disease. Among the risk factors for cardiovascular disease are high blood pressure, obesity, diabetes, stress, smoking, age, family history of heart disease, high cholesterol, and insufficient physical activity.[3]. Lifestyle modifications including stopping smoking, losing weight, exercising, and managing stress might reduce some of these risk factors. Medical history, physical examination, and imaging tests including electrocardiograms, echocardiograms, cardiac MRIs, and blood tests are used to *diagnose heart disease*. Lifestyle adjustments, drugs, medical treatments like angioplasty coronary artery bypass surgery, or implanted devices like pacemakers or defibrillators can *treat heart disease* [4]. Through the help of the enormous volumes of patient data that are readily available due to the increasing number of modern healthcare systems, it is now feasible to create prediction models for cardiac disease (also known as Big Data in Electronic Health Record Systems). Large datasets are analyzed from multiple perspectives using machine learning, which is thought of as a data-sorting technique that turns the findings into concrete knowledge [5].

This research is to develop a novel machine learning method that can accurately categorize a number of high-definition datasets and then assess how well it performs in contrast to other excellent models. The study provided the following important contributions:

1) This study's usage of a private HD dataset is one of its main contributions. Between 2022 and 2024, 200 data samples were freely submitted by Egyptian specialty hospitals. From these people, we were able to collect about thirteen features.
2) This investigation addresses the pressing need for early HD prediction in Saudi Arabia and Egypt, where The HD rate is rising quickly. The authors created a mobile application for the real-time prediction of heart disease by applying machine learning (ML) classification algorithms on a combined dataset that included both private and CHDD records.
3) This work makes an important contribution by combining XGBoost and a semi-supervised model. This method predicts HD accurately using a combined dataset. It is a new method compared to earlier studies. The research's stated goal was to predict HD using the combined datasets and the SF-2 feature subset. The following rates were achieved: 97.57% for accuracy, 96.61% for sensitivity, 90.48% for specificity, 95.00% for precision, 92.68% for F1 score, and 98% for AUC.
4) Using SHAP techniques, an explainable artificial intelligence strategy has been constructed to comprehend how the system anticipates its results. The use of SMOTE to increase the overall number of balanced cases in the dataset is of additional importance to this study. The proposed technique is trained on a balanced dataset using SMOTE to increase the performance of heart disease prediction.
5) It is also important for this study to use SMOTE to increase the total number of balanced cases in the dataset. To improve the performance of heart disease prediction, the suggested method is trained using SMOTE on a balanced dataset.
6) The ML techniques applied in this article were addi- tionally optimized with hyperparameters. We have tuned the hyperparameters for all the ML classifiers. The proposed method got 97.57% accuracy rates with hyperparameters that were optimized when the combined datasets and the SF-2 feature subset were used.
7) Additionally, to identify the classifier that achieves the most accurate HD prediction rate, the study assessed 10 distinct ML classification algorithms. The XGBoost technique was identified as a highly accurate classifier to predict HD after assessing the performance of ten algorithms. The proposed app's capacity for adaptability is shown by applying a domain adaptation method. This shows the ability of the proposed approach to be implemented in various environments and communities, in addition to the initial datasets used in this article. All things considered, this work presents fresh concepts and methods that greatly progress the field of ML-based HD prediction systems. The results of the study may be advantageous to the healthcare industries linked to the prevalence of heart disease in Saudi Arabia and Egypt.

II RELATED WORK AND COMPARATIVE STUDY

Heart disease is a leading cause of death worldwide. Predicting its possibility accurately can aid in averting it. It has been demonstrated that ML algorithms can

accurately forecast heart conditions based on a variety of medical data inputs. An overview of recent and earlier studies that have used machine learning algorithms to forecast heart disorders is provided in this section. ML techniques like as SVM, ANN, DT, LR, and RF have been used in a number of research to evaluate medical data and forecast heart conditions.

A recent study by [6] used models of ML to predict the risk of cardiac disease in a multi-ethnic population. The authors utilized a large dataset of electronic health record data and linked it with socio-demographic information to stratify CVD risks. The models achieved high accuracy in predicting CVD risk in the multi-ethnic population. Similarly, another study by [7] applied a deep learning (DL) algorithm to predict coronary artery disease (CAD). To train the DL model, the researchers used coronary computed tomography angiography (CCTA) images and clinical data. High accuracy in predicting the presence of CAD was attained by the model that was provided. A study by [8] utilized different models of ML for predicting CVD depending on clinical data. The models used by the researchers included DTs, K-nearest neighbor (KNN), and RFs. The authors reported high accuracy in predicting CVD using these models. Likewise, a study by [9] used ML techniques to determine what factors contribute to heart disease risk. The authors utilized the National Health and Nutrition Examination Survey (NHANES) data to determine risk fac- tors related to coronary heart disease. The authors reported that the proposed ML algorithm was effective in identifying risk factors. Another research study by [10] investigated different ML algorithms' accomplishments in predicting heart diseases. The authors used several models, including ANN, DT, and LR. The authors reported that the models achieved high accuracy in predicting heart diseases.

ML algorithms are now commonly used to forecast heart conditions and have demonstrated good accuracy in a number of research. ML algorithms have been used to forecast various cardiac disorders, including CAD and CVD, by taking into account medical data factors such as clinical data, sociodemographic information, and medical pictures. The papers we studied have demonstrated the effectiveness of models such as DTs, DL, ANN, RF, and KNN in predicting cardiac disorders. It is anticipated that more suitable models and characteristics will be created for precise cardiac disease prediction as machine learning algorithms continue to evolve.

Previous studies on HD prediction have shown that MLapproaches may effectively recognize features linked to the disease and build trustworthy prediction models. However, more work is needed to close these gaps in the body of current knowledge. Here are some gaps and how the proposed approach fills them.

- HD prediction research has employed one ML method, such as DT, LR, RF, or SVM. Although each of these algorithms has demonstrated potential, no thorough evaluation or comparison of ML techniques exists. This limits generalizability and complicates the search for the optimal HD predictor. This gap is filled by the suggested study. It compares and evaluates 10 ML classifiers including Naive Bayes, SVM, voting, XGBoost, AdaBoost, bagging, DT, KNN, RF, and LR. Using performance measures like accuracy, sensitivity, precision, specificity, F1-score, and AUC, the article evaluates which algorithm is the best in terms of HD prediction.
- The reason Unbalanced classes in HD prediction datasets make it difficult to make accurate predictions for the minority class (HD-positive patients). Although some studies have attempted to address this issue by using oversampling or undersampling, a thorough analysis of the techniques and their effects on prediction accuracy is required. The proposed article closes this gap by addressing the issue of unequal classes as well. SMOTE is used to guarantee that the dataset is balanced. This study looks at how well SMOTE works to improve HD prediction accuracy and how it affects the effectiveness of various ML algorithms.
- There is a demand in the literature for practical apps that can self-diagnose and detect HD. Mobile applications and other solutions have been recommended, but their efficacy, usability, and applicability to varied datasets and demographics need additional study. The proposed paper develops a smartphone app that allows users to enter HD-related symptoms for rapid predictions to fill this gap. Usability, accessibility, and adaptability to varied datasets and demographics are the app's goals. Domain adaptation is utilized to evaluate the proposed system's flexibility and ensure its real-world effective- ness. The research

article aims to improve HD research and early diagnosis and prevention in high-prevalence countries like India and America by addressing these gaps.

III ML CLASSIFICATION TECHNIQUES FOR PREDICTION

On an assortment of datasets, ML classification approaches have been extensively employed to predict CVD. The purpose of this section is to apply 10 ML classifiers to extract key features that improve CVD prediction, and talk about the recent and earlier studies on ML classification techniques for prediction.

- *Logistic Regression*: It is a well-known method that machine learning uses to classify CVD predictions. When applied to a dataset of 735 patients, LR produced a greater accuracy of 87.63% for CVD prediction in the study by [11]. In a related work, [12] employed LR to predict CVD on a dataset of 3980 patients, with an accuracy of 70.44%. In order to estimate the risk of CAD in females, the study of [13] used LR and achieved a sensitivity of 70%.
- *Random Forest*: RF is another popular technique for classification using ML. In the research by [14], RF achieved an accuracy of 76.90% for predicting CVD in a dataset of 847 patients.
- *K-Nearest Neighbor*: KNN is another algorithm that predicts CVD. In the research conducted by [16], KNN achieved an accuracy of 80.40% on a dataset of 303 patients. A similar study by [17] used the KNN technique for predicting the risk of CVD, and it achieved an accuracy of 85.76%.
- *Decision Tree*: A classifier used to forecast the risk of CVD is called a DT. Using a dataset of 4231 patients, the DT in the study by [18] obtained an accuracy of 79.3%. In a different study by [19], the DT algorithm was utilized to predict the likelihood of cardiac events with an accuracy of 85.75% on a dataset of 303 individuals.
- *Bagging*: It is a technique of ensemble learning that couples many models to improve classification accuracy. A study done by [20] used the bagging algorithm to predict the risk of CVD. It obtained an accuracy of 89.9% on a dataset of 303 patients.
- *Adaptive Boosting* (AdaBoost): It is an algorithm of ensemble learning that couples many weak classifiers to produce one strong classifier. A study by [16] used AdaBoost for predicting CVD, and it achieved an accuracy of 73.60% with a dataset of 303 patients.
- *eXtreme Gradient Boosting* (XGBoost): It is another technique of ensemble learning that couples many models to improve accuracy. A study done by [23] used the XGBoost algorithm for predicting the risk of heart diseases. It achieved an accuracy of 87.50% on a dataset of 303 patients.
- *Support Vector Machine*: SVMis a strong technique

  A study by [22] used for classification and regression used the SVM algorithm for predicting the risk of CAD and obtained an accuracy of 85.7% on a dataset of 445 patients.
- *Naive Bayes*: It is a classification algorithm that is probabilistic. An accuracy of 50% was achieved by a study by [13] that used Naive Bayes to predict the risk of CAD in females.

ML classification techniques have been widely used for predicting CVD. The ten classifiers discussed in this section have shown promising results in detecting the risk of CVD. LR, RF, and KNN algorithms have shown high accuracy in classifying the risk of CVD. Ensemble learning techniques, such as bagging, AdaBoost, and voting, have improved the classification accuracy compared to single classifiers. The accuracy of CVD risk prediction can be enhanced by employing several ML classifiers. Further research can be conducted in this area to enhance the forecast and diagnosis of CVD.

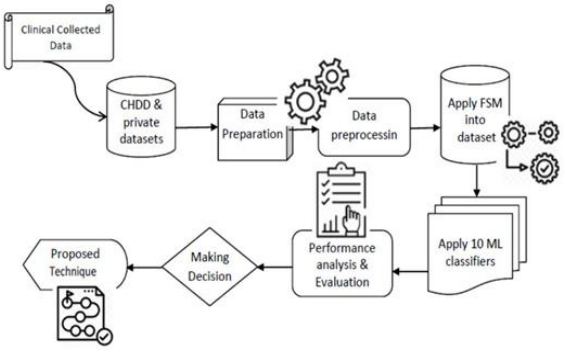

FIGURE 1. The proposed approach sequences for heart disease prediction

IV THE PROPOSED HEART DISEASE PREDICTION APP

This section explains how the suggested ML app for the prediction of cardiac ailments was implemented, including the methodology and ML algorithms employed. The sequences for predicting cardiac problems in the suggested system are displayed in Figure 1. The first step was to collect and preprocess the dataset in order to eliminate any necessary discrepancies (e.g., null occurrences needed to be replaced with average values). The dataset was separated into two distinct groups, each of which was called the training dataset and the test dataset. After that, a number of different classification algorithms were implemented in order to determine which one offered the best accuracy with regard to these datasets. Information may be included in these cloud-stored shared data. The Electronic Health Records (EHRs) stored and shared in the cloud usually contain patients' sensitive information (patient's name, telephone number and ID number, etc.) and the hospital's sensitive information (hospital's name, etc.). If these EHRs are directly uploaded to the cloud to be shared for research purposes, the sensitive information of patient and hospital will be inevitably exposed to the cloud and the researchers. Besides, the integrity of the EHRs needs to be guaranteed due to the existence of human errors and software/hardware failures in the cloud. Therefore, it is important to accomplish remote data integrity auditing on the condition that the sensitive information of shared data is protected.

A. THE PROPOSED METHODOLOGY

Naive Bayes, SVM, voting, XGBoost, AdaBoost, bagging, DT, KNN, RF, and LR classifiers are the ML techniques that are investigated in this study. These algorithms can aid doctors and data analysts in making correct diagnoses of heart condition. This article includes journals, recent research, published publications, and recent data on cardiovascular disease. The methodology outlined in [1] provides a framework for the proposed model. The approach is a series of actions that convert unprocessed data into patterns that may be consumed and recognized. As illustrated in Figure 1, the suggested method is divided into three phases: data collecting in the first stage, feature value extraction in the second stage, and data exploration in the third stage. Data preprocessing addresses missing values, data cleaning, and normalization, depending on the procedures used [2]. The ten classifiers (A1, A2,..., A10) were then used to classify the pre-processed data. Lastly, we used a variety of performance metrics to assess the correctness and performance of the proposed model once it was implemented. In this model, a Reliable Prediction System for Heart Disease (RPSHD) was created using a range of classifiers. Age, sex, blood pressure, cholesterol, and electrocardiogram are among the 13 medical parameters that this model employs to make predictions [3].

B. DATASETS AND DATASET FEATURES

In order to forecast heart disease, this study uses both the CHDD and a private dataset. There are 200 samples in the private dataset compared to 303 in the CHDD dataset, and both share the same characteristics. There are 503 records in the merged dataset, and each one has 13 attributes (such as laboratory, clinical, and demo-graphic parameters). Numerous features in the datasets can be utilized to predict cardiac disease, including age, gender, blood pressure, cholesterol levels, electrocardiogram readings-ECG, chest pain, exercise-induced angina, blood sugar with fasting condition, max heart rate achieved, oldpeak, coronary artery, thalassemia, and other clinical and laboratory measurements, as shown in Table 2.The outcome variable known as ''Target'' takes a binary value and refers to the heart disease predicting feature (i.e., it indicates whether or not cardiac disease is present).

Figure 2 shows the percentage distribution of individuals with heart disease in the combined datasets. A total of 503 samples have been gathered, and 45.9% of those have been diagnosed with HD, while the remaining 54.1% of individuals have not been infected with the disease.

An efficient visualization method for comprehending data distribution and spotting possible outliers is the boxplot. One can gain insight into the distribution of various HD-related features or variables by applying boxplots to an HD dataset. Figure 3 shows the boxplots for the HD dataset. The distribution of scores for HD detection is shown in this figure using boxplots. There was an abnormality in each of the graphs we acquired. Eliminating them will lower the data's median, which may make it more difficult to reliably identify HD. However, this approach has more advantages than the others; it may save lives by

detecting heart disease infection early, when medical treatment is most effective.

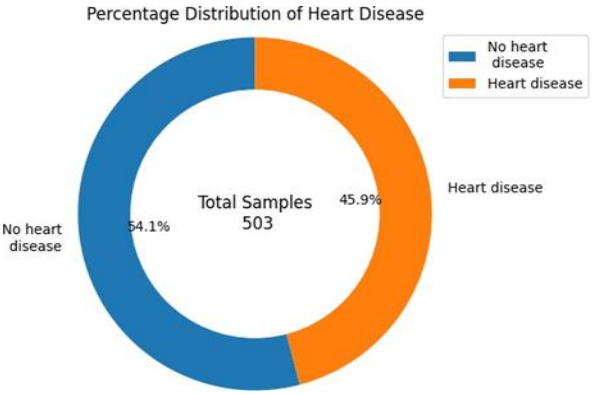

FIGURE 2. The percentage distribution of heart disease in the combined dataset.

C. DATASETS PREPARATION

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 orderly and well-written enhanced dataset was selected by taking into account the patient's history of cardiac issues as well as other medical issues.

The dataset studied in this research is a combination of accessible public WBCD and chosen private datasets. Partitioning the two datasets in this way allows us to use the method of holdout validation. 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 method. Greater dependence and information collecting are shown by larger numbers. The significance of features offers important information about each feature's applicability and capacity for prediction in a dataset. As shown in Figure 4, the thalach feature receives the highest value of 13.65% using this reciprocal information technique, while the fbs feature receives the lowest importance of 1.91%.

D. FEATURE SELECTION

In this research, we perform feature selection and classification using the *Scikit-learn* module of Python [20]. Initially, the processed dataset was analyzed using several different ML classifiers, including RF, LR, KNN, bagging, DT, AdaBoost, XGBoost, SVM, voting, and Naive Bayes, which were evaluated for their overall accuracy. In the second step, we used the Seaborn libraries from Python to create heat maps of correlation matrices and other visualizations of correlations between different sets of data. Thirdly, a wide variety of *feature selection methods* (FSM) such as analysis of variance (ANOVA), chi-square, and mutual information (MI) were applied. These strategies are explained and are indicated by the acronyms FSM1, FSM2, and FSM3, respectively. Finally, the performance of several algorithms was compared for the identified features. The validity of the analysis was demonstrated using *accuracy*, *specificity, precision, sensitivity*, and *F1 score*. The *StandardScaler* method was used to standardize every feature before it passed into the algorithms.

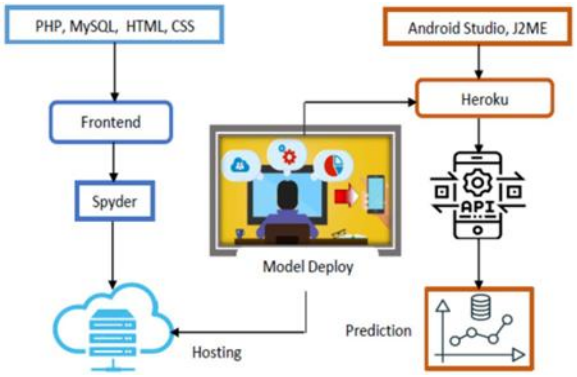

FIGURE 3. The ML-based HD prediction app design process

E.THE OUTCOME OF DIFFERENT FEATURE SELECTION METHODS

The feature weights and the ANOVA F value technique are used to calculate the F value for every pair of characteristics. The results of the ANOVA F test are shown in Table 4(a). The RES, CM, and FBS elements contribute the least to the score, whereas the EIA, CPT, and OP features are the most significant. Another method for figuring out how closely each attribute connects to the goal is chi-square. In this method, the first three features that are the most significant are MHR, OP, and CMV, whereas TS, REC, and FBS, respectively, are the least important ones. The MI technique is utilized in FSM3. To evaluate the degree of mutual dependency between

features, this approach calculates the mutual information between them. A score of 0 indicates complete independence between the two features under consideration; a larger num- ber indicates a greater dependence CPT, TS, and CMV are the three features that are most dependent on each other in this case, whereas FBS and REC are the features that are independent of each other. important factors that can be utilized for predicting the probability of having heart disease. Furthermore, REC, FBS, RBP, and CM all have lower total scores across all three FSMs. Because of all these features, three distinct groups are chosen to be included depending on their score. SF-1, SF-2, and SF-3 were the abbreviations that were given to each of the three different sets of features, respectively.

F.PROPOSED APP DEPLOYMENT

Using ML algorithms and HD symptoms, the suggested method was included into a mobile app framework to make real-time HD predictions. J2ME, PHP, HTML, MySQL, CSS, XML, and Android Studio were used in the implementation of the suggested application.

The XGBoost classifier with SMOTE using the combined datasets and SF-2 feature subset was chosen based on the research's assessment of performance criteria. Various integrated development environments (IDEs) have been used to deploy the model, including Spyder and Python IDEs. In addition, we implemented an Android app to demonstrate the prediction system's capabilities in real time and evaluate its functionality. Android Studio was used for developing the user interface of this application. The Java programming language was our primary language for coding. To implement the model, we added the Pickle package to Android Studio. Finally, we used Heroku to host the API for the proposed application. The process framework diagram for the proposed app to predict HD using ML is shown in Figure 3. Both the web-based app and the mobile app, which constitute the proposed app, have been deployed [26].

V.EXPERIMENTAL RESULTS AND ANALYSIS

Jupyter Notebook is used to predict cardiac illnesses from a dataset. It makes it easier to visualize the dataset's many data relation graphs and makes document development, including live coding, easier. The Pandas and NumPy libraries in Python are used to clean the CHDD in the first stage of this study. The dataset is then preprocessed using Python's Scikit-learn module's Standard Scaler function.

In the second step of the process, each feature's importance is calculated using a feature selection approach, and then three sets of features (SF) are generated. Thirdly, the dataset was separated into training and testing sets. A total of 75% of the data is utilized for training, while the other 25% is utilized for testing. Finally, ten distinct ML algorithms were trained using this 75% of test data. For the aim of predicting heart disease, the method with the best performance was selected.

A. PERFORMANCE EVALUATION

The authors assess and describe the performance of the suggested system in this subsection. Accuracy, sensitivity, specificity, and F1-score were among the evaluation criteria used to compare the performances of various algorithms. True positive (TP), true negative (TN), false positive (FP), and false negative (FN) data were used to assess these performance metrics. These measures are the subject of the following subsection. The algorithm with the best outcomes is given after this evaluation. The confusion matrix can be used to assess the effectiveness of a classification model, as shown in Figure 4.

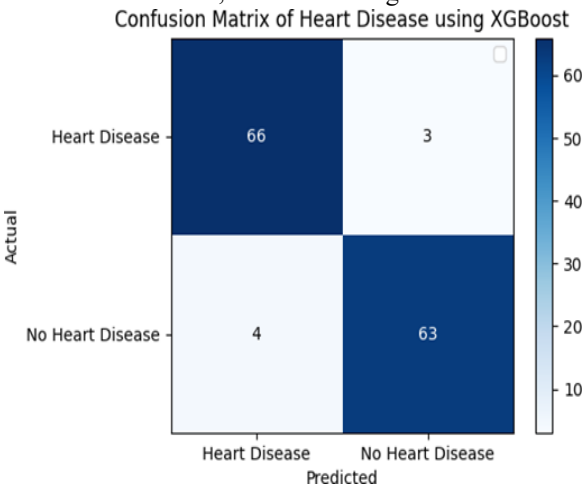

FIGURE 4. Confusion matrix of the HD dataset using XGBoost and SMOTE

Figure 4 illustrates the predicted values of TP, FP, TN, and FN for the XGBoost classifier using SMOTE. Each element in this confusion matrix represents the number of cases for both the actual classes and the

predicted classes that have a particular set of labels. As an illustration, the matrix has a total of 63 cases (TP) of heart disease classifications, 3 cases (FP) of diagnosis classified as ‘‘heart disease’’, 4 cases (FN) of diagnosis classed as ‘‘no heart disease’’, and 66 cases (TN) of distinct ‘‘heart disease’’ classifications.

VI. LIMITATIONS

Although the suggested method for HD prediction utilizing a mobile app that uses machine learning has demonstrated promising results and has the potential to be used, there are number of drawbacks that must be noted:

1. Quality and availability of datasets: The availability and quality of testing and training datasets have an impact on the effectiveness and dependability of machine learning models. In our investigation, we used private datasets and Cleveland Heart Disease. Data quality, representativeness, and availability might all be constrained. Applying the suggested method to a larger sample with a range of extra sources may be challenging due to this constraint.
2. *Imbalanced classes*: The effectiveness of ML classifiers might be affected by the presence of unbalanced classes, where one class is considerably more common than the other. The researchers in this study used SMOTE to address this issue. Although SMOTE aids in class balance, it is not an optimal solution and may produce synthetic minority class samples. This restriction may cause predictions to be biased and less accurate when applied to real-life situations.
3. *Mobile app acceptance and usability*: One major contribution to this study is the development of a mobile-based app that allows users to input symptoms and get real-time HD prediction. Engagement and adoption of the mobile app by users are crucial to the success of the proposed technique. Therefore, to guarantee application performance in a real- world setting, future work must assess important factors like user experience, privacy concerns, and accessibility.

VII. CONCLUSION AND FUTURE WORK

In order to identify the most significant features that are highly helpful in predicting heart disease, we employed a variety of feature selection techniques during the research study. Ten distinct machine learning techniques with SMOTE were then applied to the features that had been chosen. Each algorithm produced a distinct score using a separate set of characteristics. Three methods were used to choose features: ANOVA, chi-square, and MI. These methods were applied to three selected feature groups, namely, SF-1, SF-2, and SF-3, respectively. The best model and feature subset were determined using ten ML classifiers. The classifiers used were Naive Bayes, SVM, voting, XGBoost, AdaBoost, bagging, DT, KNN, RF, and LR. A well-known open-access dataset and numerous cross-validation processes were employed to evaluate the suggested algorithms and measure the heart disease detection system’s performance accuracy. When compared to all other algorithms, the performance of XGBoost was more significant. The XGBoost classifier performed best with the SF-2 feature subset, with 97.64% accuracy, 96.61% sensitivity, 90.48% specificity, 95.00% precision, a 92.68% F1 score, and a 98% AUC.

The study used a domain adaption technique to show that the suggested system is flexible. Through the introduction of novel ideas and methods, this work has significantly advanced the field of ML-based HD prediction applications. The diagnosis and prognosis of HD in India and America may benefit from these findings. Lastly, people may quickly and correctly anticipate cardiac disease by entering symptoms into a smartphone app. In conclusion, the best XGBoost approach is used by a mobile app to forecast heart disease. Among other things, we suggest collecting more personal information from more patients in order to get more precise In clinical scenarios, the use of explainable models and interpretable features is not only beneficial but mandatory. Explainable Artificial Intelligence (XAI) methods have shown to increase the performance of models by providing transparency and fostering trust among clinicians and patients.

The adoption of XAI methods addresses several critical aspects:

- Ethical and Legal Issues
- Verification with Clinical Literature
- Acceptance and Trust

To further develop the useful application of AI in healthcare, future research will continue to concentrate on enhancing model interpretability and coordinating AI predictions with clinical knowledge. By using extra strategies like LIME (local interpretable model-agnostic explanations) and consulting with clinical specialists to guarantee the interpretations, we hope to further improve the explainability of our models conform to medical practice and expertise. We hope to close the gap between sophisticated machine learning models and their real-world implementation in clinical settings by implementing these explainable AI approaches, which will ultimately lead to more transparent, dependable, and efficient healthcare solutions.

In order to make our study more comprehensive, we must highlight the advantages of explainable AI techniques and talk about how we intend to use these processes in our upcoming research. The main advantages and our suggested future paths are as follows:

Benefits of Explainable AI Methods:

- Enhanced Transparency and Interpretability
- Ethical and Legal Compliance
- Improved Clinical Outcomes

In summary, our heart disease prediction model will be much more transparent, dependable, and well-liked by patients and clinicians if explainable AI techniques are incorporated into it. The creation of strong explanatory mechanisms to aid in clinical decision-making and enhance patient outcomes will be a top priority in our upcoming work.

REFERENCE

[1] HOSAMF.EL-SOFANY "Predicting_Heart_Diseases_Using_Machine_Learning_and_Different_Data_Classification_Techniques " date of publication 1 August 2024, date of current version 12 August 2024.Digital Object Identifier 10.1109/ACCESS.2024.3437181

[2] 2023). *World Health Organization. Cardiovascular Diseases (CVDs)*. Accessed: May 5, 2023. [Online]. Available: https://www.afro.who.int/ health-topics/cardiovascular-diseases.

[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. Liu, X. Sun, Y. Liu, X. Yang, Y. Xu, and X. Sun, ‘‘Deep learning- based prediction of coronary artery disease with CT angiography,’’ Jpn.

[7] N. Zakria, A. Raza, F. Liaquat, and S. G. Khawaja, ‘‘Machine learning based analysis of cardiovascular disease prediction,’’ J. Med. Syst., vol. 41, no. 12, p. 207, 2017.

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

[9] 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.

[10] A. Shoukat, S. Arshad, N. Ali, and G. Murtaza, ‘‘Prediction of cardiovascular diseases using machine learning: A systematic review,’’ J. Med. Syst., vol. 44, no. 8, p. 162, Aug. 2020.

[11] G. R. Shankar, K. Chandrasekaran, and K. S. Babu, ‘‘An Analysis of the Potential Use of Machine Learning in Cardiovascular Disease Prediction,’’

[12] N. Khandadash, E. Ababneh, and M. Al-

Qudah, ‘‘Predicting the risk of coronary artery disease in women using machine learning techniques,’’ J. Med. Syst., vol. 45, p. 62, Apr. 2021.

[13] S. Moon, W. Lee, and J. Hwang, ‘‘Applying machine learning to predict cardiovascular diseases,’’ Healthcare Inform. Res., vol. 25, no. 2, pp. 79–86, Jun. 2019.

[14] M. Lakshmi and A. Ayeshamariyam, ‘‘Machine learning techniques for prediction of cardiovascular risk,’’ Int. J. Adv. Sci. Technol., vol. 30, no. 3,pp. 11913–11921, Mar. 2021.

[15] M. R. Hassan, S. Huda, M. M. Hassan, J. Abawajy, A. Alsanad, and G. Fortino, ‘‘Early detection of cardiovascular autonomic neuropathy: A multi-class classification model based on feature selection and deep learning feature fusion,’’ Inf. Fusion, vol. 77, pp. 70–80, Jan. 2022.

[16] S. P. Mikles, H. Suh, J. A. Kientz, and A. M. Turner, ‘‘The use of model constructs to design collaborative health information technologies: A case study to support child development,’’ J. Biomed. Informat., vol. 86, pp. 167–174, Dec. 2018.

[17] M. R. Delavar, M. Motwani, and M. Sarrafzadeh, ‘‘A comparative study on feature selection and classification methods for cardiovascular disease diagnosis,’’ J. Med. Syst., vol. 39, no. 9, p. 98, Sep. 2015.

[18] C. Puelz, S. Acosta, B. Rivière, D. J. Penny, K. M. Brady, and C. G. Rusin, ‘‘A computational study of the Fontan circulation with fenestration or hepatic vein exclusion,’’ Comput. Biol. Med., vol. 89, pp. 405–418, Feb. 2017.

[19] Q. Z. Mirza, F. A. Siddiqui, and S. R. Naqvi, ‘‘The risk prediction of cardiac events using a decision tree algorithm,’’ Pakistan J. Med. Sci., vol. 36, no. 2, pp. 85–89, Mar./Apr. 2020.

[20] 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.

[21] S. Jhajhria and R. Kumar, ‘‘Predicting the risk of cardiovascular dis- eases using ensemble learning approaches,’’ Soft Comput., vol. 24, no. 7, pp. 4691–4705, Jul. 2020.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] S. Akbar, R. Tariq, and A. Basharat, ‘‘Heart disease prediction using dif- ferent machine learning approaches: A critical review,’’ J. Ambient Intell. Humanized Comput., vol. 11, no. 5, pp. 1973–1984, 2020.

[27] 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.

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] F. Pedregosa, G. Varoquaux, and A. Gramfort, ‘‘Scikit-learn: Machine learning in Python,’’ J. Mach. Learn. Res., vol. 12, pp. 2825–2830, Oct. 2011.

[35] H. El-Sofany, S. A. El-Seoud, O. H. Karam, Y. M. Abd El-Latif, and I.

[36] A. T. F. Taj-Eddin, ‘‘A proposed technique using machine learning for the prediction of diabetes disease through a mobile app,’’ Int. J. Intell. Syst., vol. 2024, pp. 1–13, Jan. 2024.