

Received 1 July 2024, accepted 29 July 2024, date of publication 1 August 2024, date of current version 12 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3437181

RESEARCH ARTICLE

Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques

HOSAM F. EL-SOFANY^{ID}

College of Computer Science, King Khalid University, Abha, Saudi Arabia
Cairo Higher Institute for Engineering, Computer Science and Management, Cairo, Egypt
e-mail: helsofany@kku.edu.sa

This work was supported by the Deanship of Research and Graduate Studies at King Khalid University under Grant RGP1/129/45.

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. Death rates can be lowered by the early detection of heart diseases and the constant monitoring of patients by physicians. Unfortunately, heart disease cannot always be detected accurately, and a doctor cannot be in touch with a patient 24/7. Machine learning (ML) has the potential to aid in diagnostics by providing a more precise basis for prediction and making decisions using data given by healthcare sectors throughout the world. This study aims to employ several feature selection methods to develop an accurate ML technique for heart disease prediction in its earliest stages. The feature selection process was performed using three distinct methods, namely, chi-square, analysis of variance (ANOVA), and mutual information (MI). 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. The experimental findings demonstrated that the XGBoost classifier achieved the optimal performance using the combined datasets and SF-2 feature subset with the following rates: 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. The development of an explainable artificial intelligence approach that makes use of SHAP methodologies is being done to get an understanding of how the system predicts its ultimate results. 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 Cardiovascular disease, heart disease, machine learning app, ML algorithms, SDG 3, SHAP, SMOTE.

The associate editor coordinating the review of this manuscript and approving it for publication was Fahmi Khalifa^{ID}.

I. INTRODUCTION

The *heart* is a muscular organ that represents the central pumping organ of the circulatory system. It is responsible for pumping blood throughout the body and is one of the components of the cardiovascular system. The term

“cardiovascular system” can also refer to the system of arteries, veins, and capillaries that carry blood throughout the body. Several distinct forms of heart illness, which collectively fall under the umbrella term *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].

As stated in Sustainable Development Goal (SDG) 3 of the United Nations, each person should be healthy and happy, this research investigates cardiovascular disease. Heart disease is often diagnosed by observing the patient’s symptoms and conducting a physical examination. Some of the *risk factors* for cardiovascular disease include smoking, age, heart disease history in the family, high cholesterol levels, lack of time spent in physical activities, high blood pressure, obesity, diabetes, and stress [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].

It is now possible to construct prediction models for heart disease with the assistance of the vast amounts of patient data that are easily accessible as a result of the growing number of recent healthcare systems (also known as Big Data in Electronic Health Record Systems). Machine learning is considered a data-sorting approach that analyzes large datasets from various viewpoints and then transforms the results into tangible knowledge [5].

The objective of the study is to provide an ML approach for heart disease prediction. ML algorithms were evaluated on large, open-access heart disease prediction datasets. Finally, the most accurate and dependable algorithm was chosen as the final model for an Android mobile app. This study aims to construct an innovative machine learning technique that is capable of properly classifying several high-definition datasets and then evaluate its performance in comparison to that of other first-rate models. The study provided the following important contributions:

- 1) One of the key contributions of this research is the use of a private HD dataset. Egyptian specialized hospitals voluntarily provided 200 data samples between the years 2022 and 2024. We were able to gather around 13 features from these participants.
- 2) This work deals with the immediate requirement for early HD prediction in Egypt and Saudi Arabia, where

the HD rate is rapidly increasing. Through the application of ML classification algorithms to a combined dataset consisting of both CHDD and private datasets, the authors developed a mobile-based app for the instantaneous prediction of heart disease.

- 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) To understand how the system predicts its outcomes, an explainable artificial intelligence approach utilizing SHAP methodologies has been developed.
- 5) 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.
- 6) The ML techniques applied in this article were additionally 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.

Overall, this work introduces novel ideas and techniques that significantly advance the field of ML-based HD prediction systems. The healthcare sectors that are associated with heart disease incidences in Egypt and Saudi Arabia may both benefit from the research’s findings.

II. RELATED WORK AND COMPARATIVE STUDY

A major death cause globally is heart disease. Accurate prediction of its likelihood can help in preventing it. ML Algorithms have been proven to predict heart diseases effectively based on various medical data parameters. This section presents a review of current and previous research that has utilized ML algorithms to predict heart diseases. Several studies have utilized ML algorithms like SVM, artificial neural network (ANN), DT, LR, and RF to analyze medical data and predict heart diseases.

A recent study by [6] used models of ML to predict the risk of cardiac disease in a multi-ethnic population. The authors

TABLE 1. Comparative study of using ML classifiers to predict heart diseases.

Year	Authors	Datasets used	Algorithms used (ML classifiers)	No. of classifiers	Accuracy obtained
2021	Liu et al. [24]	UCI heart disease	LR, RF, KNN, SVM, Naive Bayes	5	93%
2020	Hussein et al. [25]	Cleveland heart disease	LR, KNN, DT	4	84%
2020	Akbar et al. [26]	Cleveland heart disease	RF, SVM, Naive Bayes	3	87%
2019	Zarshenas et al. [27]	Cleveland heart disease	XGBoost, DT, SVM, Naive Bayes	4	91%
2019	Kaur and Singh [28]	UCI heart disease	AdaBoost, DT, KNN, RF, LR	5	97%
2018	Li et al. [29]	Cleveland heart disease	Voting, Bagging, RF, SVM, Naive Bayes	5	90%
2018	Zhang et al. [30]	Cleveland heart disease	AdaBoost, DT, RF, KNN, LR, SVM, Naive Bayes	7	92%
2017	Wu et al. [31]	Cleveland heart disease	RF, SVM, Naive Bayes	3	87%
2016	Ahmed et al. [32]	Cleveland heart disease	LR, KNN, DT	3	77%
2007	Chen et al. [33]	Cleveland heart disease	LR, KNN, DT	3	85%
2024	Proposed technique	Cleveland heart disease, and private datasets	Naive Bayes, SVM, Voting, XGBoost, AdaBoost, Bagging, DT, KNN, RF, LR	10	97.57%

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). The researchers utilized clinical data and coronary computed tomography angiography (CCTA) images to train the DL model. The presented model achieved high accuracy in predicting the presence of CAD. 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 factors 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 have become widely accepted in predicting heart diseases and have shown high accuracies in various

studies. Considering medical data parameters like clinical data, socio-demographic information, and medical images, ML algorithms have been utilized to predict different heart diseases such as CAD and CVD. The studies we have reviewed have showcased those models like DTs, DL, ANN, RF, and KNN can effectively predict heart diseases. With the increasing advancements in ML algorithms, it is expected that more appropriate models and features will be developed for accurate heart disease prediction.

Previous studies on HD prediction have shown that ML approaches 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 used one ML algorithm, such as DT, LR, RF, or SVM. Each of these algorithms has shown promise, but there is no comprehensive comparison or assessment of ML approaches. This restricts generalizability and makes it difficult to find the best HD predictor. The proposed study addresses this gap. 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.

- Accurate prediction is challenging for the minority class (HD-positive patients) due to imbalanced classes in HD prediction datasets. While some research has tried to solve this problem by employing oversampling or undersampling, an extensive evaluation of the methods and how they affect prediction accuracy is necessary. The imbalanced classes issue is also addressed in the proposed article, which eliminates this gap. To ensure that the dataset is balanced, SMOTE is used. The effectiveness of SMOTE in enhancing the accuracy of HD predictions and its effects on the efficiency of different ML algorithms are examined in this work.
- 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 effectiveness. The research article aims to improve HD research and early diagnosis and prevention in high-prevalence countries like Egypt and Saudi Arabia by addressing these gaps.

A. COMPARATIVE STUDY OF HD PREDICTION APPLYING ML CLASSIFIERS

ML is a powerful tool for predicting HD. It has the potential to enhance patient outcomes by its ability to facilitate early detection and personalized treatment. This section introduces a comparative analysis of heart disease prediction using ten ML classifiers, including Naive Bayes, SVM, voting, XGBoost, AdaBoost, bagging, KNN, DT, RF, and LR (see Table 1).

The results indicated that ML classifiers could improve heart disease prediction accurately, with the highest achieved being 97% by [28] using AdaBoost, DT, RF, KNN, and LR on the UCI dataset. Several studies utilized the Cleveland heart disease dataset (CHDD), with accuracies ranging from 77% [32] to 92% [30] using various ML algorithms such as AdaBoost, DT, RF, KNN, LR, SVM, and Naive Bayes. Hence, ML classifiers could improve the certainty of heart disease forecasting, enabling early detection and personalized treatment. Nonetheless, more investigation is essential to validate these classifiers' accuracy using larger datasets, to increase the generalizability and reproducibility of the results.

III. ML CLASSIFICATION TECHNIQUES FOR PREDICTION

The classification techniques of ML have been widely used for predicting CVD on various datasets. This section aims to

discuss the current and previous research on ML classification techniques for prediction and apply ten ML classifiers to extract essential features that enhance CVD prediction.

- **Logistic Regression:** It is a famous technique used by ML for the classification of CVD prediction. In the research conducted by [11], LR was used on a dataset of 735 patients, which achieved higher accuracy for CVD prediction with 87.63%. A similar study conducted by [12] used LR for CVD prediction on a dataset of 3980 patients, and it achieved 70.44% accuracy. The study of [13] employed LR for predicting the risk of CAD in females and obtained 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. Using leave-one-out cross-validation, research by [15] demonstrated an algorithm that could detect the early or unusual phases of cardiovascular autonomic neuropathy (CAN) with an AUC score of 0.931.
- **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 DT is a classifier used for predicting the risk of CVD. In the research by [18], the DT achieved an accuracy of 79.3% on a dataset of 4231 patients. In another study by [19], the algorithm of the DT was used for predicting cardiac event risk possibility with an accuracy of 85.75% on a dataset of 303 patients.
- **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.
- **Voting:** It is a technique of ensemble learning that couples many models to produce the final decision for classification. A study by [21] used voting for predicting CVD, and it achieved an accuracy of 92.20% with a dataset of 303 patients.
- **Support Vector Machine:** SVM is a strong technique used for classification and regression. A study by [22]

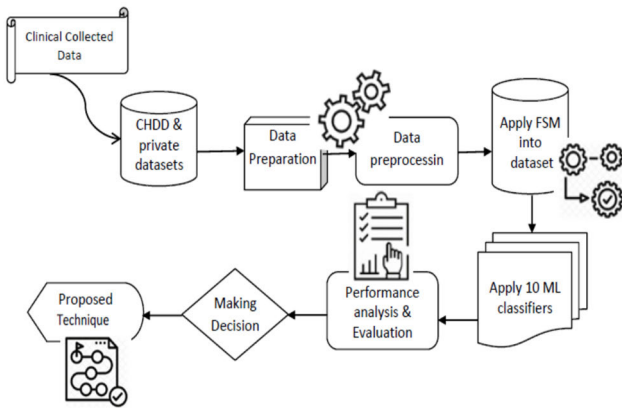


FIGURE 1. The proposed approach sequences for heart disease prediction.

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 probabilistic algorithm used for classification. A study by [13] employed Naive Bayes for predicting the risk of CAD in females and obtained an accuracy of 50%.

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.

IV. THE PROPOSED HEART DISEASE PREDICTION APP

In this section, we explain the approach used and the ML algorithms applied in implementing the proposed ML app for the prediction of cardiac illnesses. Figure 1 shows the proposed system's sequences for predicting heart diseases. To begin with, the dataset was required to be gathered and preprocessed so that any necessary inconsistencies could be removed from it (e.g., null occurrences needed to be replaced with average values). The dataset was divided into two distinct groups, which were referred to as the test dataset and the training dataset, respectively. Following that, several distinct classification algorithms were put into action to identify the one that provided the highest level of accuracy concerning these datasets.

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

cardiac disease. Recent data on cardiovascular illness as well as journals, recent research, and published publications are all part of this article. A framework for the suggested model is provided by the methodology as in [1]. The methodology is a set of steps that transforms raw data into consumable and identifiable data patterns. The proposed approach consists of three stages: the first stage is data collection; the second stage extracts specific feature values; and the third stage is data exploration, as shown in Figure 1. Depending on the procedures employed, data preprocessing deals with the missing values, cleansing of the data, and normalization [2]. The data that underwent pre-processing were then classified using the ten classifiers (A1, A2, ..., A10). Finally, after putting the suggested model into practice, we evaluated its performance and accuracy using a range of performance measures. Using a variety of classifiers, a Reliable Prediction System for Heart Disease (RPSHD) was developed in this model. This model uses 13 medical factors for prediction, among which are age, sex, cholesterol, blood pressure, and electrocardiograph [3].

B. DATASETS AND DATASET FEATURES

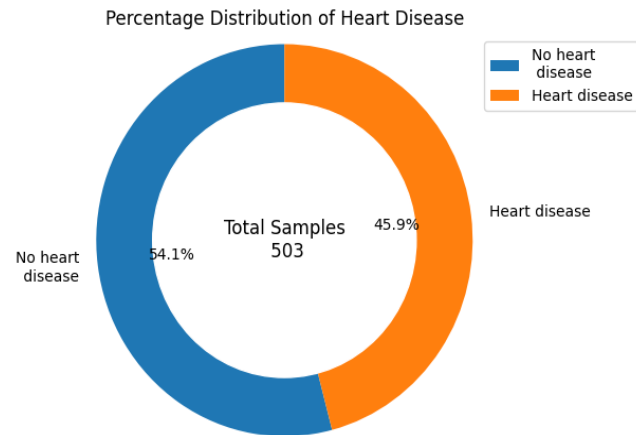
This research employs both the CHDD and a private dataset for heart disease prediction. The CHDD dataset has 303 samples, while the private dataset has 200, and they have the same features. The combined dataset contains 503 records, and 13 features are associated with each one (including demographic, clinical, and laboratory parameters). The datasets have many features that can be used for heart disease prediction 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.

Boxplots are an effective visualization technique for understanding the distribution of data and identifying potential outliers. By applying boxplots to a dataset related to HD, one can get insights into the distribution of a variety of HD-related features or variables. The HD dataset's boxplots are illustrated in Figure 3. Boxplots are used to illustrate the distribution of scores for HD detection in this figure. Every graph we obtained had an anomaly. Removing them will cause the median of the data to drop, which might make it harder to detect HD accurately. On the other hand, this method offers more benefits than the others; by identifying heart disease infection at an early stage, when medical care is most beneficial, this diagnostic could preserve lives.

TABLE 2. The used features from the CHDD.

Feature no.	Feature name	Feature code	Description	Values type
1	Age	AGE	Age of patient	Number of years
2	Gender	GEN	Patient sex	Female = 0, male = 1
3	Chol	CHOL	Evaluation of a patient's cholesterol levels	mg/dl
4	Trestbps	BRP	Blood resting pressure	Mm
5	CP	CPT	Chest pain types	Typical angina = 1, atypical angina= 2, nonanginal pain = 3, asymptomatic = 4
6	Fbs	FBS	Blood sugar in fasting case	< or > 120 mg/dl (true = 1, false = 0)
7	Thalach	MHR	Maximum rate achieved on heart	Continuous
8	RestEcg	REC	Electrocardiograph by resting	0 = no abnormalities, 1 = normal, 2 = left ventricular hypertrophy (possible or certain)
9	Oldpeak	OP	ST depression when compared to rest taken quantity	Continuous
10	Exang	EIA	Angina caused by exercise	1 = there is pain, 0 = there is no pain
11	Ca	CMV	Count of main vessels colored by fluoroscopy	0-3
12	Slope	PES	Peak exercise ST segment slope	Up sloping = 0, flat = 1, down = 2
13	Thal	TS	Thallium stress	Negative = 0, positive = 1, inconclusive = 2
14	Target		target variable representing diagnosis of heart disease using the angiographic disease status.	0 = no heart disease (< 50% diameter narrowing) 1 = heart disease (> 50% diameter narrowing)


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

C. DATASETS PREPARATION

In this research, preprocessing was performed on collected data. The CHDD has four inaccurate CMV records and two erroneous TS entries. Incorrect data is updated to reflect the best possible values for all fields. Then, StandardScaler is employed to normalize all the features to the relevant coefficient, ensuring each feature has a zero mean and one variance. By considering the patient's history of cardiac problems and following other medical concerns, an organized and composed augmented dataset was chosen.

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

holdout validation method. In this study, 25% of the data is in the test dataset, compared to 75% in the training dataset. The mutual information method is used in this research to measure the interdependence of variables. Larger numbers indicate greater dependency and information gathering. The importance of features provides valuable insights into the relevance and predictive power of each feature in a dataset. Using this reciprocal information technique, the *thalach* feature is given the highest value of 13.65%, while the *fbs* feature is given the lowest importance of 1.91%, as illustrated in Figure 4.

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 in Table 3 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.

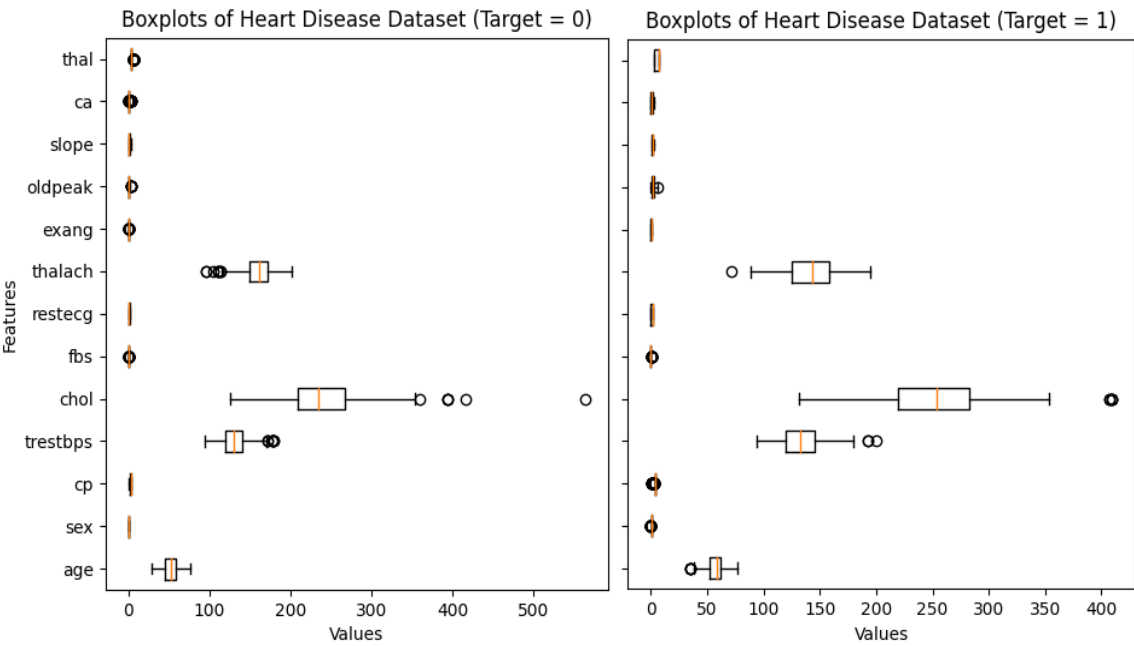


FIGURE 3. Boxplots of the combined heart disease dataset.

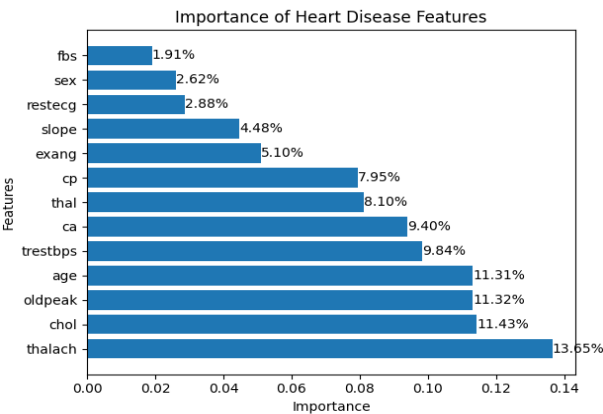


FIGURE 4. The importance of the heart disease dataset features.

E. THE OUTCOME OF DIFFERENT FEATURE SELECTION METHODS

The F value for each pair of features is determined by using the ANOVA F value technique and the feature weights. Table 4(a) presents the findings of the ANOVA F test. The EIA, CPT, and OP features provide the most importance to the score, while the RES, CM, and FBS features contribute the least. Chi-square is another approach that determines the degree to which every feature relates to the target. Table 4(b) shows the chi-square outcomes. 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

TABLE 3. The methods used to select features.

Univariate selection algorithm	Code and Description	Formula used
ANOVA F value	FSM1: The ANOVA test is a method of enhancing classification accuracy through the reduction of high-dimensional data, the identification of relevant features using feature space, and the measurement of similarity between features.	$F = \frac{\sum_{j=1}^J N_j (x_j - \bar{x})^2 / (J - 1)}{\sum_{j=1}^J ((N_j - 1) S_j^2) / (N - 1)}$
Chi-square	FSM2: To determine which of several nonnegative features is most valuable, a chi-squared score must be computed. It represents the difference between the observed and expected values.	$\chi^2 = \sum \frac{(o_j - e_j)^2}{e_j}$
Mutual information	FSM3: Mutual information is a measurement of the relationship between features.	$I(X; Y) = H(Y) - H\left(\frac{Y}{X}\right)$

between the two features under consideration; a larger number indicates a greater dependence. The MI score results are shown in Table 4(c). 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. Table 4 illustrates 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

TABLE 4. Feature score using FSM1, FSM2, and FSM3.

Feature No	Feature Code	(a) Score using FSM1		(b) Score using FSM2		(c) Score using FSM3	
		Score	Order	Score	Order	Score	Order
1	AGE	17.12	9	24.29	7	1.01	11
2	GEN	26.79	8	8.58	10	1.05	9
3	CHOL	3.20	12	24.94	6	1.08	7
4	BRP	7.46	10	15.82	8	1.03	10
5	CPT	70.77	2	63.60	4	1.17	1
6	FBS	1.24	13	1.20	13	1.00	12
7	MHR	66.12	4	189.32	1	1.10	5
8	REC	6.78	11	3.98	12	1.00	13
9	OP	69.55	3	73.64	2	1.09	6
10	EIA	71.95	1	39.91	5	1.10	4
11	CMV	65.05	5	71.89	3	1.11	3
12	PES	41.90	6	10.80	9	1.08	8
13	TS	32.80	7	6.90	11	1.14	2

TABLE 5. Three distinct feature groups (SF-1, SF-2, and SF-3).

Feature groups	Selected features
SF-1	AGE, GEN, CHOL, BRP, CPT, FBS, MHR, REC, OP, EIA, CMV, PES, TS
SF-2	AGE, GEN, CHOL, CPT, MHR, OP, EIA, CMV, PES, TS
SF-3	AGE, GEN, CPT, MHR, OP, EIA, CMV, PES, TS

that were given to each of the three different sets of features, respectively. Table 5 shows these feature sets that were selected for additional investigation.

F. PROPOSED APP DEPLOYMENT

The proposed technique was integrated into a mobile app framework using ML algorithms and HD symptoms to predict HD instantaneously on real data. We have implemented the proposed application using J2ME, PHP, HTML, MySQL, CSS, XML, and Android Studio.

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 (see Table 6). 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

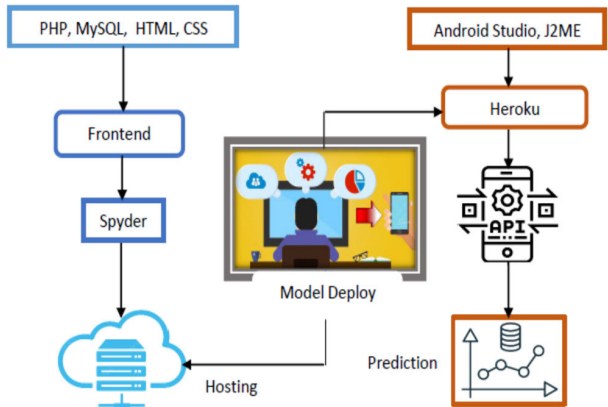


FIGURE 5. The ML-based HD prediction app design process.

Figure 5. Both the web-based app and the mobile app, which constitute the proposed app, have been deployed [26].

V. EXPERIMENTAL RESULTS AND ANALYSIS

Predicting heart diseases from a dataset is done using Jupyter Notebook. It simplifies the visualization of different data relation graphs of the dataset and facilitates the creation of documents including live coding. In the first step of this research, the CHDD is cleaned with Python’s *Pandas* and *NumPy* libraries. After that, the dataset is preprocessed with the *StandardScaler* method from Python’s *Scikit-learn* module [34]. 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

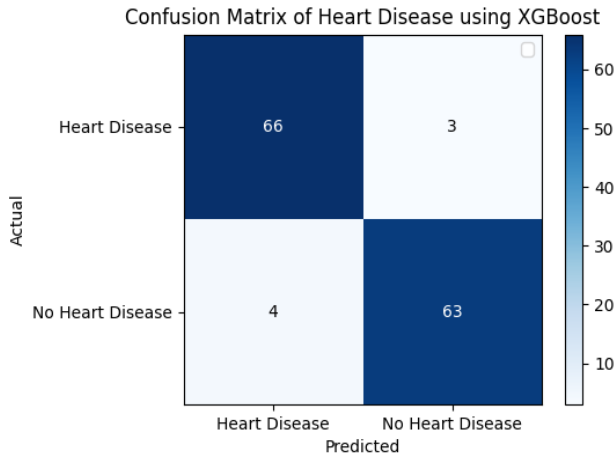


FIGURE 6. Confusion matrix of the HD dataset using XGBoost and SMOTE.

heart disease, the method with the best performance was selected [35].

A. PERFORMANCE EVALUATION

In this subsection, the authors evaluate and explain the proposed system's performance. Different algorithms and their comparative performances were presented based on evaluation metrics including accuracy, sensitivity, specificity, and F1-score. These performance measures were evaluated using true positive (TP), true negative (TN), false positive (FP), and false negative (FN) data. The next subsection focuses on these measurements. Following this evaluation, the algorithm with the greatest results is provided. Figure 6 demonstrates how the confusion matrix may be used to evaluate a classification model's performance.

Figure 6 illustrates the predicted values of T_P , F_P , T_N , and F_N 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 classified as "no heart disease", and 66 cases (TN) of distinct "heart disease" classifications.

Figure 7 presents the correlation between the important features of SF-2 using SMOTE. The y-axis values include thalach, chol, sex, age, slope, exang, oldpeak, ca, cp, and thal. Positive or negative correlation coefficients show a significant relationship between the two variables, whereas -1 and 1 indicate no association. It is essential to keep in mind that the only thing that can be detected via the use of correlation is the linear link that exists between the variables. The prediction for the patient is correlated with each of those variables at a level of at least 70% correlation.

Figures 8 and 9 show the scatter and density plots among four selected features in the SF-2 dataset. These scatter and density graphs are beneficial for exploring the relationships and distributions of variables in the HD dataset. They can

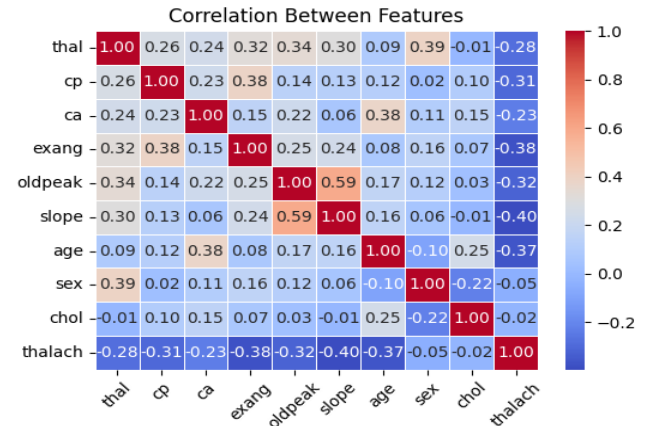


FIGURE 7. Correlation between features of SF-2 using SMOTE.

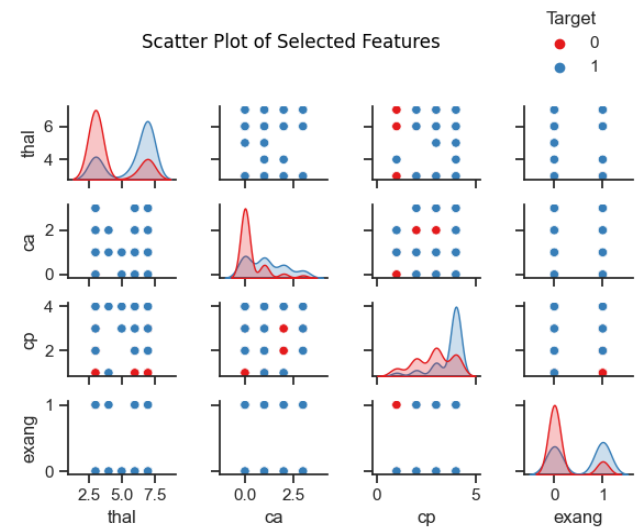


FIGURE 8. Scatter plot among four selected features in the SF-2.

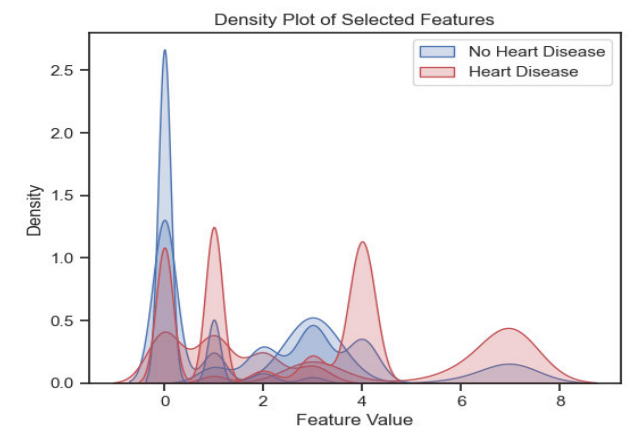


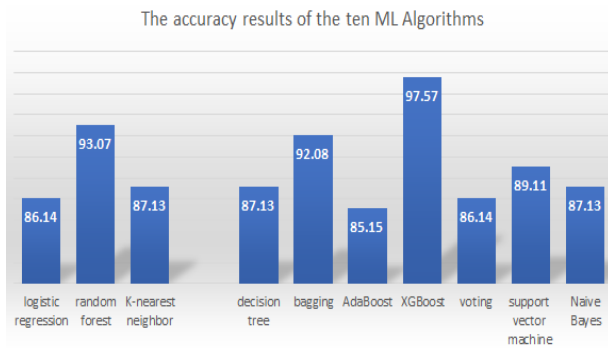
FIGURE 9. Density plot among the first four important features in the SF-2.

provide insights into correlation, concentration, outliers, and patterns that may exist among the four variables (exang, cp, ca, and thal).

- **Accuracy:** The proposed model's accuracy was developed to determine what percentage of samples has been

TABLE 6. Accuracy, sensitivity, and specificity of ML techniques using SF-1, SF-2, and SF-2.

Algo. code	ML algorithms	Accuracy %			Sensitivity %			Specificity %			F1 score		
		SF-1	SF-2	SF-3	SF-1	SF-2	SF-3	SF-1	SF-2	SF-3	SF-1	SF-2	SF-3
A1	Naive Bayes	87.13	87.13	87.13	91.53	91.53	91.53	80.95	80.95	80.95	83.95	83.95	83.95
A2	SVM	89.11	89.11	88.12	94.92	94.92	93.22	80.95	80.95	80.95	86.08	86.08	85.00
A3	Voting	88.12	86.14	86.14	94.92	94.92	94.92	78.57	73.81	73.81	84.62	81.58	81.58
A4	XGBoost	93.17	97.57	93.19	94.92	96.61	94.92	90.48	90.48	90.48	91.57	92.68	91.57
A5	AdaBoost	86.14	85.15	85.15	91.53	91.53	88.14	78.57	76.19	80.95	82.50	81.01	81.93
A6	Bagging	89.11	92.08	91.09	94.92	94.92	93.22	80.95	88.10	88.10	86.08	90.24	89.16
A7	Decision Tree	89.11	87.13	93.07	93.22	91.53	93.22	83.33	80.95	92.86	86.42	8395	91.76
A8	KNN	86.14	87.13	88.12	89.83	93.22	89.83	80.95	78.57	85.71	82.93	8354	85.71
A9	Random Forest	93.07	93.07	93.07	94.92	94.92	94.92	90.48	90.48	90.48	91.57	91.57	91.57
A10	Logistic Regression	86.14	86.14	88.12	93.22	93.22	93.22	76.19	76.19	80.95	82.05	82.05	85.00

**FIGURE 10.** The accuracy results of the ten ML algorithms.

accurately classified. Accuracy is computed using the formula given in (Eq. 1), which is based on the confusion matrices:

$$\text{Accuracy} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \quad (1)$$

- **Sensitivity (or recall):** Sensitivity measures the rate of truly positive results and implies that all values should be evaluated positively. Additionally, sensitivity is calculated as “the proportion of correctly detected positive samples”. Sensitivity is determined by the following formula:

$$\text{Sensitivity} = \frac{T_P}{T_P + F_N} \quad (2)$$

- **Specificity:** It predicts that all values will be negative and is determined by calculating the fraction of real negative situations. Specificity is determined mathematically by

$$\text{Specificity} = 1 - \left(\frac{F_P}{F_P + T_N} \right) \quad (3)$$

- **Precision:** It determines classifier accuracy and may be calculated from the information given. This is presented

by comparing real TP versus predicted TP. The formula in (Eq. 4) shows how the accuracy measure verifies the proposed method’s behavior:

$$\text{Precision} = \frac{T_P}{T_P + F_P} \quad (4)$$

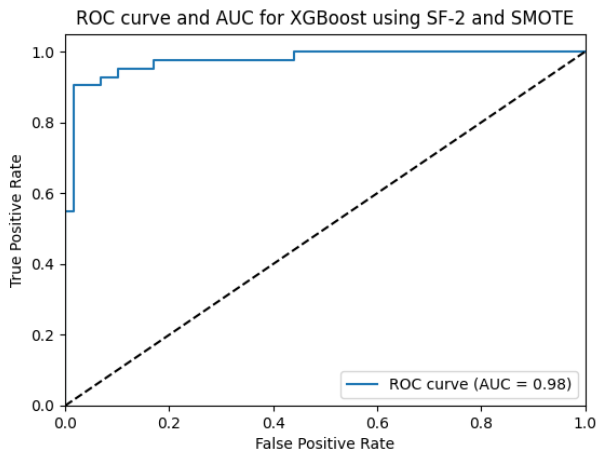
- **F-measure:** It is a statistical measure that is employed in the process of evaluating the efficacy of a classification model. It does this by determining the harmonic mean of the accuracy and recall measurements, giving each of these metrics an equal amount of weight. It enables the performance of a model to be described and compared using a single score that takes into consideration both the recall and precision of the model’s predictions and is calculated using the following formula:

$$F - \text{measure} = \frac{2 * (\text{precision} * \text{recall})}{\text{precision} + \text{recall}} \quad (5)$$

The performance of a classifier has been represented and evaluated with the use of a confusion matrix, as shown in Figure 6. T_P measures how many individuals are accurately classified into the sick positive class. The percentage of healthy people who are appropriately labeled as being in the negative class is known as T_N . The number of times that healthy persons were incorrectly diagnosed as being sick is referred to as the F_P . When the number of healthy persons is mistakenly predicted, this is known as F_N . A comparison of the various performance indicators across 10 ML algorithms is presented in Table 6. These AI classifiers were applied to the combined dataset that contained SF-1, SF-2, and SF-3 feature subsets. Based on its accuracy of 97.75%, sensitivity of 96.61%, specificity of 90.48%, precision of 95.00%, and F1 score of 92.68% for the SF-2 feature group (see Table 6), the XGBoost classifier had the best overall performance.

TABLE 7. Performance outcomes for XGBoost classifier using SF-2 and SMOTE.

Algorithm	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUC
XGBoost	97.57%	96.61%	90.48%	95.00%	92.68%	0.98

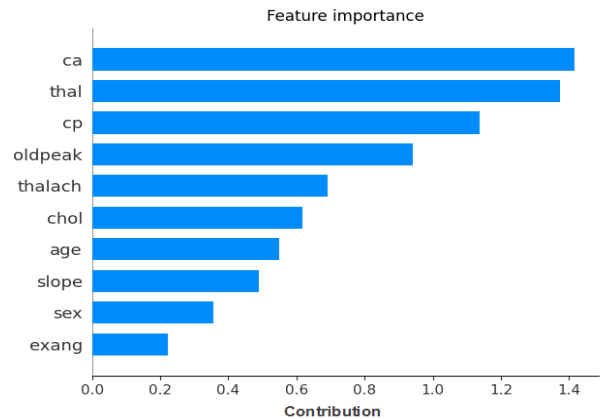
**FIGURE 11.** AUC and ROC curve for the XGBoost classifier using SMOTE.

B. EXPERIMENTAL EVALUATION OF SYSTEM PERFORMANCE

The accuracy of every technique is displayed in Table 6, along with the processed dataset that was analyzed using those algorithms. In terms of the accuracy of each technique, A4's accuracy calculation for SF-2 was the greatest accurate (97.57%), followed by its accuracy calculations for SF-1 and SF-3 (93.17% and 94.19%), respectively. A9 computed an accuracy of 93.07% over all three SFs, placing it in second place. On the other hand, A5 determined that SF-1 and SF-3 had a low accuracy of 85.15% among all classifiers. A3 and A10 likewise provided a low level of accuracy for SF-2 and SF-3, coming in at 86.14% and 86.12%, respectively. The other methods have an accuracy between 87.13% and 90.00%. Furthermore, this finding shows that the XGBoost algorithm method using the SF-2 is the most effective for processing the dataset. Figure 10 shows all the various accuracy rates that may be achieved for the ten ML techniques using all of SF-2.

In this study, all the algorithms' sensitivities were evaluated. Table 6 displays the sensitivity scores obtained from the ten ML techniques using SF-1, SF-2, and SF-3, respectively. A5's sensitivity to SF-3 was the lowest (88.14%). A8 rated both SF-1 and SF-3 (89.83% and 89.83%, respectively). A4 (XGBoost) reported the highest sensitivity for SF-2 as well, at 96.61%; A2, A3, A4, A6, and A9 reported the second-highest sensitivity, at 94.92%.

The analysis of specificity was performed on each of these techniques, and the results are summarized in Table 6. A3 scored the lowest (73.81%) for SF-2 and SF-3. A4 and A9 scored the highest (90.48%) for all SFs, based on the results of the analysis. When compared to the results of the other

**FIGURE 12.** Explainable AI interpretation of the XGBoost feature importance.

techniques, A7 for SF-3 (92.86%) provided the best score with SF-3 only.

C. DISCUSSION

In this study, a variety of ML techniques were implemented for the early recognition of CVD, and a combined dataset (CHDD and private datasets) was employed for both testing and training purposes. The ML model was then tested and trained on the source and target datasets using a domain adaptation approach. The proposed HD prediction technique used in this study was first trained using a private dataset with 200 cases. After that, the system was evaluated using the combined dataset with 503 cases. To be more specific, we employed a total of ten well-known ML algorithms including Naive Bayes, SVM, voting, XGBoost, AdaBoost, bagging, DT, KNN, RF, and LR, denoted by (A1, A2, ..., A10), each with a unique set of selected features. The values of the ANOVA F statistic, the chi-square test, and the MI statistic were the statistical methods utilized in the classification of relevant aspects that were more useful for CVD prediction. Five different evaluation standards were used to compare and rate the performance of the different ML techniques that used SMOTE: these were accuracy, sensitivity, precision, specificity, and F1 score. The outcomes of the experiment showed that algorithm A4 obtained the best rate of accuracy (97.57%) for SF-2, and the accuracy rate achieved by A9 was second best (93.07%) across all three SFs presented in Table 6. A4 likewise obtained the greatest score possible for sensitivity (96.61%), as well as the best score possible for specificity (90.48%), while testing for SF-2, as shown in Table 6. The result of the F1 score demonstrated that A4 had the highest score of 92.68% for SF-2 (see Table 6), while A9 obtained the highest score

of 91.57% for SF-1, SF-2, and SF-3, and A6 obtained the highest score of 90.24% for SF-2. Because A4 has the best performance when employed with SF-2, this method is the most reliable technique in terms of accuracy, specificity, and sensitivity. In terms of F1 score, A9 is the more accurate predictive model for all SFs, which places it as the second-best predictive algorithm overall. As a result of this research, we have concluded that provides the highest performance rate. As a consequence of this, it is permissible to conclude that XGBoost is an effective method for predicting heart diseases. When combining the results of multiple different ML algorithms, an accuracy range of 85.15 to 97.57% was achieved in the vast majority of cases. Finally, the proposed system is a mobile app employing RF ML.

Figure 8 shows the mobile app's quick heart disease detection using real data and the highest successful classification. Figure 11 illustrates the XGBoost classifier's receiver operating characteristic (ROC) curve with SF-2, which demonstrates the model's performance across all classification thresholds, with an AUC of 0.98 (see Table 7).

We have integrated the SHAP methodology into the proposed work to provide a unified measure and to interpret the contribution of each feature to the model's predictions, thus enhancing the transparency and interpretability of our machine learning approach. Here's how SHAP enhances our model:

- 1) *Feature Contribution Explanation*: SHAP values help us understand the impact of each individual feature on the model's predictions. For instance, in the case of predicting heart disease, SHAP can elucidate how features like cholesterol levels, blood pressure, and age contribute to the final prediction, allowing clinicians to see which factors are most influential.
- 2) *Individual Prediction Analysis*: We can analyze individual predictions using SHAP values to understand the reasoning behind a specific patient's prediction. This is particularly important in clinical settings, where understanding the rationale behind a prediction can guide further medical investigation or treatment.
- 3) *Global Model Insights*: SHAP not only provides local interpretability (individual predictions) but also offers global insights into the model's behavior across the entire dataset. This helps to identify which features are generally most important and how they interact with each other.
- 4) *Trust and Adoption in Clinical Practice*: To implement AI models in healthcare settings, the clarity offered by SHAP values is essential. Understanding and validating the model's decision-making process increases clinicians' trust and use of machine learning models.

In our study, after identifying the best performing model (XGBoost with the SF-2 feature subset), we applied SHAP to interpret the results. The SHAP analysis revealed which features were most important and how they affected the predictions. This information is not only valuable for model validation but also for offering actionable insights to healthcare

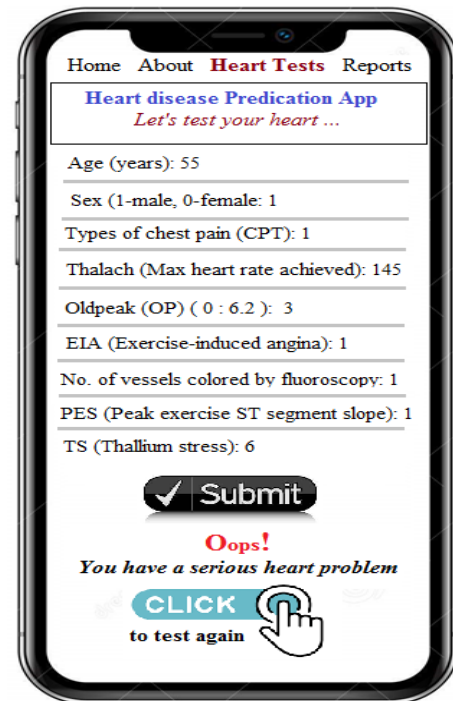


FIGURE 13. A snapshot of the proposed mobile app that predicts heart disease immediately.

professionals. Figure 12 shows an explainable AI interpretation of the importance of the features using the SHAP library and XGBoost classifier.

The last step was to implement the system into a mobile app by employing an XGBoost with SMOTE. As shown in Figure 13, the most effective classification was used in the development of the mobile app that provides an immediate and accurate diagnosis of HD using real data.

Using ML classifiers for HD prediction is the goal of this work. The experiment findings proved that the XGBoost algorithm was the most accurate percentage for predicting the occurrence of HD. The following features are classified as important for HD prediction according to the mutual information-based feature selection approach: thalach, chol, oldpeak, age, trestbps, ca, thal, cp, exang, slope, restecg, sex, and fbs. We have used the SMOTE method to optimize hyperparameters and oversample using the data that was collected. The XGBoost technique with SMOTE produced the best results. The study reached its goal of predicting HD, with the combined datasets, and the experimental results were 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.

VI. LIMITATIONS

Many limitations need to be acknowledged, even though the proposed approach for predicting HD using a mobile app that makes use of ML has shown encouraging results and has the promise to be employed:

1. *Dataset quality and availability*: The performance and reliability of ML models depend on the quality and availability of testing and training datasets. We employed Cleveland heart disease, and private databases in our study. There may be limitations in availability, representativeness, and data quality. This limitation could make it hard to apply the proposed approach to a broader sample with a variety of additional sources.

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. *Algorithm selection*: To determine the optimal algorithm for predicting HD, the researchers used a variety of ML techniques. Nonetheless, the selection of algorithms is arbitrary and may affect the outcome. Other algorithms that were not considered in this study might be able to achieve different trade-offs or greater accuracy. As a result, future research should carefully consider and evaluate the ML algorithms selected.

4. *Domain adaptation*: The use of techniques for domain adaptation demonstrated the adaptability of the proposed system. The proposed technique may still have some limitations when it comes to its ability to be applied to a variety of different populations or environments. More research is required to determine the technique's efficacy in a range of populations with different lifestyles, demographics, and healthcare systems. It is also important to fully address any potential restrictions and difficulties related to domain adaptation.

5. *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

During the process of this research study, we used a variety of methods for selecting features, after which we used ten different ML techniques with SMOTE to apply to the features that had been selected after determining the most important features that are very useful for predicting heart disease. Every algorithm generated a unique score based on a different combination of features. 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 demonstrated that the proposed system is adaptable using a domain adaptation approach. This work has made a significant contribution to the field of ML-based HD prediction applications by introducing unique insights and techniques. These findings have the possibility of aiding in the diagnosis and prediction of HD in Egypt and Saudi Arabia. Finally, a smartphone app allows users to enter symptoms and predict heart disease quickly and accurately. In conclusion, a mobile app predicts heart disease using the best XGBoost technique. We recommend gathering more private data from more patients to generate more accurate findings, among other possibilities.

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 [36]. The adoption of XAI methods addresses several critical aspects:

- 1) **Ethical and Legal Issues**: Explainability ensures that the decision-making process of AI models aligns with ethical standards and legal requirements, as highlighted in [37]. Transparent models aid in auditing and validating decisions, which is critical in a highly regulated healthcare sector.
- 2) **Verification with Clinical Literature**: The AI bases its decisions on scientific reasoning, as its features and models align with established clinical literature [38]. This alignment with clinical knowledge enhances the reliability of the model's predictions.
- 3) **Acceptance and Trust**: Transparent and understandable decision-making processes significantly enhance the acceptance and trust of AI systems in clinical practice [39]. Explainable models offer valuable insights into prediction processes, thereby facilitating their seamless integration into standard clinical workflows.

Our study concludes that incorporating explainable AI methods such as SHAP into machine learning models for heart disease prediction enhances their clarity, reliability, and acceptability in clinical settings.

Future work will continue to focus on improving model interpretability and aligning AI predictions with clinical expertise to further advance the practical application of AI in healthcare. We aim to further enhance the explainability of our models by applying additional techniques such as LIME (local interpretable model-agnostic explanations) and engaging with clinical experts to ensure the interpretations

align with medical knowledge and practice. By incorporating these explainable AI methodologies, we aim to bridge the gap between complex ML models and their practical application in clinical settings, ultimately contributing to more transparent, reliable, and effective healthcare solutions.

To increase the completeness of our study, it is essential to emphasize the benefits of explainable AI methods and discuss how we plan to implement these mechanisms in our future work. Here are the key benefits, as well as our proposed future directions:

Benefits of Explainable AI Methods:

1) Enhanced Transparency and Interpretability:

- *Clinical Decision Support*: Explainable AI methods, such as SHAP, provide clear insights into the contribution of each feature to the model's predictions. This transparency helps clinicians understand the rationale behind AI-generated predictions, making the decision-support process more robust.
- *Model Validation and Verification*: Using XAI methods, healthcare professionals can verify the model's predictions against clinical knowledge and literature. This validation process ensures that the model's decisions are based on medically sound principles, which is critical for clinical acceptance.

2) Ethical and Legal Compliance:

- *Accountability*: Explainable models facilitate accountability by providing a clear audit trail of the decision-making process. This is crucial for addressing ethical and legal concerns in healthcare and ensuring that AI systems comply with regulatory standards.
- *Patient Trust*: Understanding the prediction process fosters trust in the AI system among patients and clinicians. This trust is critical for widespread adoption of AI in clinical settings.

3) Improved Clinical Outcomes:

- *Personalized Treatment Plans*: XAI methods enable the identification of key factors influencing individual patient predictions. This can lead to more personalized and effective treatment plans tailored to the specific needs of each patient.
- *Early Intervention*: By providing detailed explanations for predictions, clinicians can identify early warning signs and intervene promptly, potentially improving patient outcomes.

Future Directions: We plan to incorporate the following explanatory mechanisms into our future work to strengthen trust and confidence in our AI application for heart disease prediction.

1) Integration of SHAP and LIME:

- *SHAP*: We will continue to use SHAP to provide both global and local explanations of model predictions. SHAP values will help us identify the most

influential features and understand their impact on individual predictions.

- *LIME*: We will explore the use of LIME to generate interpretable models around each prediction. Locally, LIME approximates the black-box model with an interpretable model, providing further insights into the prediction process.

2) Interactive Explanation Interfaces:

- We will develop interactive interfaces that allow clinicians to visualize and explore the explanations provided by SHAP and LIME. These interfaces will enable users to drill down into specific predictions, compare feature contributions, and gain a deeper understanding of the model's behavior.

3) Clinical Expert Collaboration:

- We will collaborate with clinical experts to ensure that the explanations generated by our AI models align with clinical practice and knowledge. Their feedback will be invaluable in refining our explanation mechanisms and ensuring their relevance and accuracy in a clinical context.

4) Continuous Model Monitoring and Improvement:

- We will implement continuous monitoring of our AI models to track their performance and the relevance of the explanations over time. This will include regular updates to the models and the explanation mechanisms based on new clinical data and feedback from healthcare professionals.

In conclusion, the integration of explainable AI methods into our heart disease prediction model will significantly enhance its transparency, reliability, and acceptance among clinicians and patients. In our future work, we will prioritize the development of robust explanation mechanisms to support clinical decision-making and improve patient outcomes.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest to report regarding the present study.

AVAILABILITY OF DATA AND MATERIALS

The corresponding author will share study datasets upon reasonable request.

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. Liu, X. Sun, Y. Liu, X. Yang, Y. Xu, and X. Sun, "Deep learning-based prediction of coronary artery disease with CT angiography," *Jpn. J. Radiol.*, vol. 38, no. 4, pp. 366–374, 2020.
- [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 analysis," *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," *J. Med. Syst.*, vol. 43, no. 12, p. 345, Mar. 2019.
- [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. Inform.*, 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. Jhahhria and R. Kumar, "Predicting the risk of cardiovascular diseases 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 different 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. 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.
- [36] V. Bento, M. Kohler, P. Diaz, L. Mendoza, and M. A. Pacheco, "Improving deep learning performance by using explainable artificial intelligence (XAI) approaches," *Discover Artif. Intell.*, vol. 1, no. 1, pp. 1–11, Dec. 2021.
- [37] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision making and a 'right to explanation,'" *AI Mag.*, vol. 38, no. 3, pp. 50–57, Sep. 2017.
- [38] C. Militello, F. Prinzi, G. Sollami, L. Rundo, L. La Grutta, and S. Vitabile, "CT radiomic features and clinical biomarkers for predicting coronary artery disease," *Cognit. Comput.*, vol. 15, no. 1, pp. 238–253, Jan. 2023.
- [39] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018.



HOSAM F. EL-SOFANY received the M.Sc. and Ph.D. degrees in computer science. He is currently an Associate Professor of CS with King Khalid University, Saudi Arabia, and Cairo Higher Institute for Engineering, Computer Science and Management, Egypt. He also supervised many M.Sc. and Ph.D. dissertations in computer science and information systems. He has published approximately 80 research papers in international refereed journals and conferences. His research interests include cloud computing, e-learning, m-learning, u-learning, fuzzy logic, cloud security, cybersecurity, and chronic disease prediction techniques using ML and DL algorithms. He reviews many international journals and conferences.

...