IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 May 2024, accepted 3 June 2024, date of publication 11 June 2024, date of current version 28 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3412077

RESEARCH ARTICLE

A Hybrid Feature Selection and Ensemble Stacked Learning Models on Multi-Variant CVD Datasets for Effective Classification

ABHIGYA MAHAJAN[1], BAIJNATH KAUSHIK[1], MOHAMMAD KHALID IMAM RAHMANI[2], (Senior Member, IEEE), AND ABDULBASID S. BANGA[2]

[1]School of Computer Science and Engineering, Shri Mata Vaishno Devi University, Katra, Jammu and Kashmir 182320, India
[2]College of Computing and Informatics, Saudi Electronic University, Riyadh 11673, Saudi Arabia

Corresponding authors: Mohammad Khalid Imam Rahmani (m.rahmani@seu.edu.sa), Abdulbasid S. Banga (a.banga@seu.edu.sa), and Baijnath Kaushik (Baijnath.kaushik@smvdu.ac.in)

ABSTRACT Predicting cardiac or heart disease has emerged as a formidable challenge in the medical domain recently. It is recognized as a major global health concern, and stands as one of the primary causes of mortality, posing a significant threat to human life. Early detection of heart disease helps to reduce mortality. This study has experimented with three benchmark datasets such as UCI Heart Disease, Framingham, and Z-Alizadeh Saini containing important clinical information for cardiac vascular disease (CVD). These three datasets' multi-variant (categorical and continuous) features, variable dimensions, and multicollinearity characteristics provide substantial challenges for machine learning (ML) and other models aiming to achieve the desired results. This study proposes a statistical feature selection (SFS) stacking framework using four feature engineering techniques, Chi-Square, Gini Index, Information Gain, and ANOVA F-test, to select the optimal features from the datasets. Further, the likelihood of developing CVD based on characteristics extracted from the three benchmark datasets using a reduced set of optimized features from the initial feature set is fed to ensemble stacked learning models: stacking using Support Vector Machine (SFS-SVM) and stacking using Cross-Validation Classifier (SFS-SCVC). The SFS-SCVC model has achieved significant performance metrics and outperformed the SFS-SVM and traditional ML models on all three datasets.

INDEX TERMS Multi-variant CVD datasets, gini index, chi-square, ANOVA F-test, stacking SVM, stacking cross-validation.

I. INTRODUCTION

CVD is a cluster of diseases involving the heart and blood vessels supplying blood to various organs and systems of the body including diseases like coronary heart disease, stroke, angina, heart failure, peripheral vascular disease, aortic dissection, etc. [1]. According to recent statistics, about one person dies every minute from heart disease. It has multifactorial causation that is, it is caused by a lot of different major risk factors including both controllable and non-controllable as depicted in Table 1.

Earlier it was more prevalent in developed countries and now its incidence is also increasing in developing countries not only the elderly but young people are also presenting with heart ailments and even deaths due to heart attacks [2], [3] making it one of the most common death-causing disease worldwide. However, its prevalence cannot be estimated accurately because many cases in developing and under-developed countries aren't even present in healthcare facilities and hospitals. Also, different countries use different methods of reporting the prevalence. So, predicting the prevalence worldwide is almost impossible [4].

Every year, a staggering number of lives, as reported by the World Health Organization (WHO), succumb to

The associate editor coordinating the review of this manuscript and approving it for publication was Gyorgy Eigner.

TABLE 1. Major risk factors behind CVD.

Controllable Factors	Non-controllable Factors
Obesity	Age
Stress	Gender
Eating habits	Genetics
Sedentary lifestyle	Family history
Smoking	Ethnicity

CVD - a striking 17.9 million individuals. Out of these heart-wrenching statistics, more than 80% of deaths are attributed to heart attacks and strokes, while a disheartening one-third of these losses occur prematurely, before reaching the age of 70. Adding to this reality is the fact that over 75% of CVD fatalities transpire in low and middle-income countries, where limited healthcare resources restrict access to comprehensive screening and predominantly focus on managing existing cases [5], [6], [7]. As a result, people are detected very late in the course of the disease and so even young productive populations are dying because of heart disease as well.

Proper and relevant history taking, clinical examination, and evaluation using ECG, ECHO cardiography, Chest X-ray, Blood work, etc. are required for accurately diagnosing heart disease [8], [9]. Many attributes have been observed in the patient's medical records and from different clinical reports. However, some of these attributes may not be relevant in the same way and others might need to be left out. Moreover, a diagnosis is less accurate if you apply all aspects simultaneously. The two dominating steps in predicting cardiovascular diseases are choosing the most important variables ignoring less important ones and selecting an appropriate classifier. Recently, notably in the medical area, technologies based on ML have enhanced our quality of life [10], [11]. Time and infrastructure are needed to diagnose accurately. Time is a crucial factor when trying to save human life. So, for early detection and treatment, the power of advancements in technology like Artificial Intelligence (AI) can be utilized in the healthcare sector [12], [13], [14]. ML is a subset of AI used to design, implement, and evaluate algorithms that enable computers to gain insights from experience and make better predictions on data.

ML is used for the diagnosis and prediction in many research papers whether the patient has been diagnosed with CVD. The results have shown an improvement in the proposed classification framework by improving the accuracy of the existing method. Finally, the key step to improve the accuracy of diagnosis in CVD is the selection of features. For example, the doctor could decide to treat a patient based on classification based on selected characteristics. In previous studies, improvements and developments in classification methods emphasize more than the best features. This study aims to investigate how different SFS techniques might enhance CVD prediction. To overcome the problem of variable dimensionality and multicollinearity in features among datasets different SFS methods have been used. CVD prediction may be possible using a novel approach that conducts thorough evaluations on benchmark datasets with and without feature selection (FS) to evaluate the relevance of characteristics, which would be beneficial to the medical community [15], [16].

A. MAJOR CONTRIBUTIONS AND HIGHLIGHTS

The significant contributions of this manuscript are mentioned below:

- Three major benchmark datasets with multi-variant features and variable dimensions were used in this proposed work. Further, standard pre-processing techniques such as handling missing values, and min-max normalization were applied.
- The SFS Classification Framework has been proposed with two stacking approaches which implement six major ML models as the base models and one as a meta-model.
- Four different statistical techniques have been used Information Gain, Gini Index, Chi-Square, and Anova F-test for handling categorical and continuous predictors to select the union of rank-based optimal feature set.
- In the first approach, Hybrid feature selection (HFS) with a Support Vector Machine (SVM) stacking framework named SFS-SVM is used where SVM is the meta classifier.
- In the second approach, HFS with a Stacking CV Classifier framework named SFS-SCVC is used where stackingCV is the meta classifier.
- Finally, the proposed framework performance and individual ML models' performance have been compared and tested on full feature space and optimal features in performance evaluation metrics like accuracy, recall, precision, and F1-score.

The succeeding sections of this manuscript are structured as follows: In the upcoming section, we delve into the studies relevant to FS techniques, CVD diagnosis, and prediction processes, highlighting their significant contributions. Section III provides a comprehensive overview of the proposed hybrid approach, encompassing details about the dataset, the ML models employed in the study, and the SFS methods utilized for feature optimization. Section IV delves into the implementation specifics, encompassing evaluation parameters in this manuscript, along with the reported proposed framework outcomes. Finally, in Section V, we summarize the main conclusions drawn from this research manuscript.

II. BACKGROUND STUDY

This section comprehensively reviews recent developments in CVD detection using ML and SFS techniques. The authors consider the advancements made in recent years, exploring innovative methodologies for effective CVD prediction. By reviewing the literature that used ML algorithms and FS techniques, we have valuable information about the progression of the research for potential CVD prediction.

Deepika and Balaji [17] combined the Grey Wolf Firefly (GF) algorithm and Differential Evolution (DE) called GF-DE, which employed effective FS using the Grey wolf (GW) optimization with the Firefly algorithm for hyperparameters tuning using the DE algorithm and the comparison have been made with the proposed system using the Cleveland and Statlog datasets based on precision, recall, F1-score and accuracy.

Zhang et al. [18] proposed a CVD prediction model by combining the embedded FS method with deep neural networks (DNN) based on the LinearSVC algorithm and L1 norm. Lasso as a penalty term was used to create a sparse weight matrix to filter out variables associated closely with CVD, and the Linear SVC algorithm was employed in the FS module after data preprocessing. They compared three weight initialization techniques: Xavier, He Normal, and Random Normal, and observed that the He initialization method yielded the best outcomes in the prediction model for heart disease. To validate their proposed model, the researchers conducted tests using a CVD dataset obtained from Kaggle. The dataset consisted of 1025 patient records from different age groups with a gender distribution of 713 males and 312 females. By computing metrics of F1-score, accuracy, recall, and precision they assessed the performance of the research work.

Gárate Escamila et al. [19] proposed a dimensionality reduction method to identify attributes associated with heart disease through FS techniques. The dataset was acquired from the UCI ML Repository, specifically the heart disease dataset comprising 74 features with their corresponding labels. By combining Chi-square and principal component analysis called CHI-PCA, in conjunction with the random forests (RF), exceptional accuracy rates were attained. The experimental results demonstrated that the performance of most classifiers improved when using CHI-PCA. However, the study's main limitation is the small sample size, limiting the generalizability of the findings.

Enhanced evolutionary FS technique combined with an ensemble model named GA-LDA (Genetic Algorithm - Linear Discriminant Analysis) was proposed by Jothi Prakash and Karthikeyan [20]. Their approach achieved a maximum accuracy of 93.65% for the Statlog dataset. Furthermore, it achieved an accuracy of 82.81% for the SPECTF dataset and 84.95% for the coronary heart disease dataset. They also showcased the performance of the proposed approach through ROC curve analysis against state-of-the-art ML methods.

The utilization of ensemble algorithms- stacking, bagging, boosting, and majority voting, was implemented by Latha and Jeeva [21]. Bagging improved the accuracy by a maximum of 6.92% while boosting improved it by a maximum of 5.94%. The majority voting with weak classifiers and stacking yielded accuracy improvements of up to 7.26% and 6.93% respectively. The majority voting excelled using a reduced feature set.

A stacking learning approach coupled with seven classifiers, including Naive Bayes, random forest, KNN, and four others, was experimented with by Książek et al. [22] with and without FS on the hepatocellular carcinoma disease of 165 HCC patient's dataset collected from Coimbra's Hospital and University Centre (CHUC). They employed normalization techniques to fill in the missing variables and the KNN algorithm to train and assess the model using organized cross-validation approaches. Their model achieved an overall 90% accuracy and 88.57% of the F1-score.

A deep CNN framework and a fuzzy c-means neural network (FNN) were proposed by Venkatesan et al. [23] for feature extraction to improve prediction accuracy and investigate cardiac complaints. The FNN categorizes sensor data to recognize the heart's condition, and evaluation performances indicate that FNN performs well in predicting cardiac complaints. Their model achieved an accuracy rate of 86.4%, outperforming other approaches.

Spencer et al. [24] investigated the comparative efficacy of eight classification models and various FS methods on four heart disease datasets taken from the UCI ML repository with 14 features and 720 instances. The Chi-squared FS technique with the BayesNet classifier achieved the highest accuracy of 85.0%.

A combination of LSTM and Angle Transform (AT) methods has been proposed by Kaya et al. [25], for the prediction of arrhythmia and congestive heart failure. The AT method utilizes angular information of neighboring signals to classify ECG signals, generating new signals ranging between 0 and 359. LSTM utilizes histograms of these signals to distinguish between ARR, CHF, and normal sinus rhythm (NSR). The proposed approach achieves a high ECG signal classification rate of 98.97% tested on MIT-BIH and BIDMC databases. Demir et al. [26] propose a method using the ECG dataset where co-occurrence matrices were created and used to extract Heralick features and achieved a success rate of 93.41% using SVM for classification.

Rahman et al. [27] proposed a self-attention-based transformer model to predict cardiovascular disease risk. It combined self-attention mechanisms and transformer networks to capture contextual information effectively. The model's interpretability was highlighted as it assigned attention weights to input components, aiding in understanding predictions. Tested on the Cleveland dataset, it achieved an accuracy of 96.51%.

The Gradient Squirrel Search Algorithm-Deep Maxout Network (GSSA-DMN) was developed by Balasubramaniam et al. [28] employing data pre-processing- log scaling and FS using Relief. The DMN, trained by GSSA, combined Gradient Descent Optimization with the Squirrel Search Algorithm. GSSA-DMN achieved accuracy, sensitivity, and specificity values of approximately 93.2%, 93%, and 91.5%, respectively, surpassing existing methods by margins of 6.97%, 5.79%, 4.50%, 3.43%, and 1.93% respectively. This indicates its superior performance in heart disease detection.

A stacking ensemble learning model, incorporating deep neural networks with a tenfold cross-validation framework was employed by Gupta et al. [29]. The model was trained using heart-related data obtained from individuals who have survived COVID-19. The proposed work achieved an accuracy of 93.23% and was also compared with baseline learning algorithms while predicting heart disease.

A. RESEARCH GAPS

Researchers have diligently focused on employing diverse FS methods and traditional classification techniques for improved prediction accuracy. They recognize that the features embedded within datasets wield a significant influence over the accuracy of predictions and the computational complexity of the ML process. Consequently, selecting the ideal subset of features during feature extraction emerges as a pivotal element in any ML endeavor. While researchers have introduced various FS techniques and classification algorithms suited to specific datasets, the FS process needs further refinement.

B. MOTIVATION

This refinement aims to find out the precise subset of features that yield dependable predictions with enhanced accuracy. The FS techniques must be able to identify the most essential features, minimizing the size of the feature set while preserving their ability to predict CVD cases accurately. Obtaining more efficiency requires optimizing computational resources and streamlining the prediction process. Hence, the quest for a novel framework capable of analyzing datasets, integrating the most effective SFS techniques, and ensuring reliable predictions with minimal features and reduced computational complexity remains an important research task.

The proposed method addresses the limitations of traditional CVD prediction methods which need improvement of their suboptimal performance and computational inefficiency. Existing approaches mainly utilize simple FS techniques and conventional classification algorithms, which leads to inadequate prediction accuracy and increased computational complexity. These methods also fail to effectively exploit the wide landscape of CVD datasets overlooking crucial features, and yielding unreliable predictions. Moreover, the dependence on outdated methods restricts the adaptability to the dynamic nature of CVD diagnosis. This results in hindrance to progress in the field. On the other hand, the proposed method exploits the advanced SFS techniques and hybrid stacking classifiers enhancing the accuracy of CVD predictions with efficient utilization of computational resources. The proposed method meticulously refines the performance of the models through fine-tuning and effective evaluation, surpassing the limitations of traditional approaches and providing the opportunity for better precision and efficiency in CVD prediction.

III. METHODOLOGY

The proposed framework improves classification accuracy by reducing the range of features available in three benchmark datasets of CVD, as illustrated in Fig. 1. The framework is designed using essential elements- data collection, data preprocessing, FS techniques, training of hybrid stacking classifiers, and model evaluation. These elements work effectively to ensure an effective system. The framework starts by gathering the data, which is processed and prepared for analysis. SFS techniques are applied to identify the most relevant and informative attributes. The hybrid stacking classifiers are trained using the selected features, combining the strengths of different models. Finally, the model's performance is evaluated to assess its effectiveness and accuracy. The stages of the proposed framework are organized and discussed further in the following subsections.

The study focuses on HFS and classification using four SFS techniques applied to the initial feature set of each dataset to get the reduced number of features generated based on scores for each dataset. A new group of features named the optimal feature set is generated by applying the union operation on all the reduced feature sets which contain a non-repeated unique set of features. After identifying the optimal attributes, distinct training, and test sets are created from the datasets. The samples are divided such that 70% of the total samples are allocated for classification and performance evaluation in the training set, while the rest is utilized for testing.

TABLE 2. Dataset summary.

Datasets	Sources	Descriptions
Dataset 1	Framingham Dataset	15 features 1-Target variable Categorical: 7 Continuous: 8
Dataset 2	UCI Heart Disease Dataset (Combination of Cleveland, Hungary, Switzerland, and the VA Long Beach databases)	13 features 1-Target variable Categorical: 6 Continuous: 7
Dataset 3	The Z-Alizadeh Saini Dataset	53 Features 1-Target variable Categorical: 34 Continuous: 19

A. DATASET DESCRIPTION

This study makes optimal use of three datasets namely Framingham, UCI ML Repository heart disease dataset (obtained from Hungary, Cleveland, Switzerland, and VA Long Beach databases), and the Z-Alizadeh Saini are shown in Table 2.

Dataset 1: The Framingham Heart Disease dataset is derived from a comprehensive longitudinal cardiovascular cohort study [30]. This dataset is specifically curated to study heart disease and includes medical, laboratory, and questionnaire data collected from 4240 participants. Among the 15 variables, 8 are numerical, representing continuous measurements, while the remaining 7 are categorical variables. Since the category variables are binary, they can have one of two alternative instances. This selection of characteristics was carefully chosen to record crucial information regarding CVD

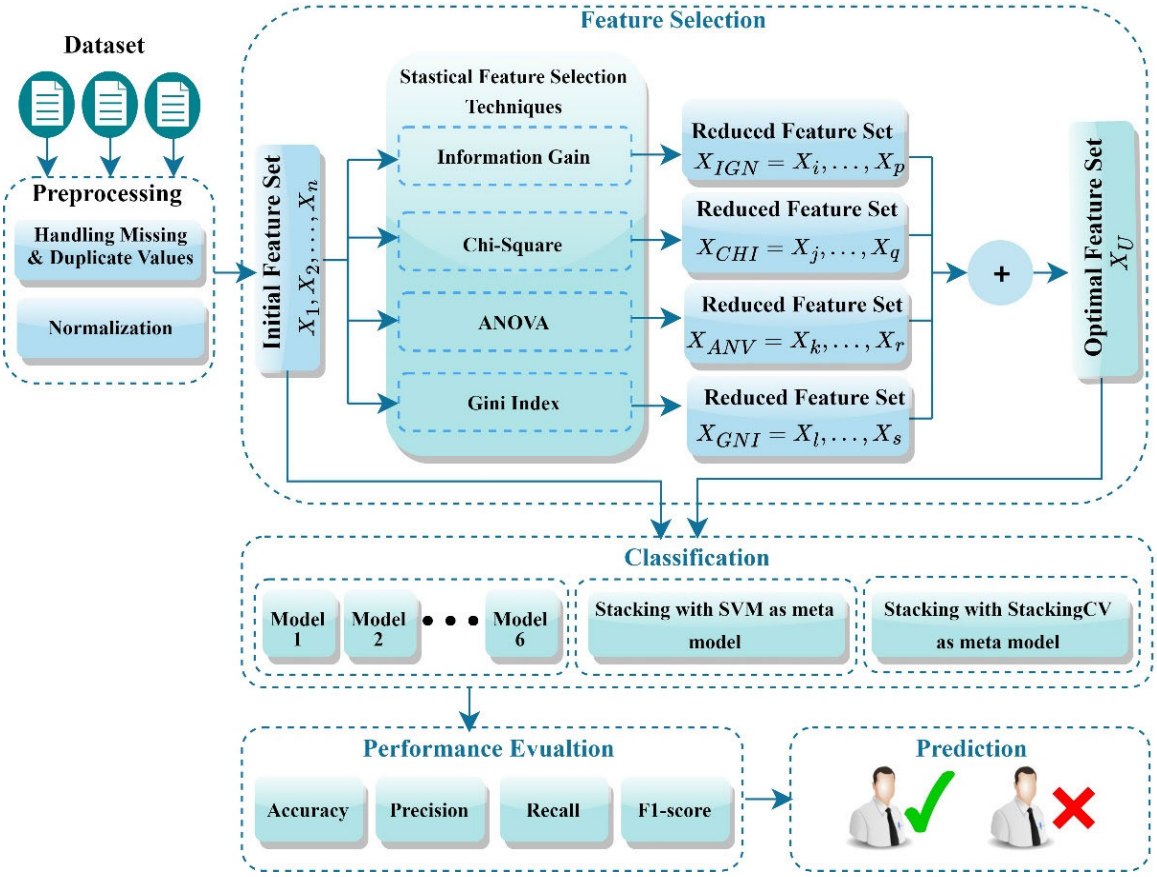

FIGURE 1. Statistical feature selection stacking framework.

and associated risk factors. Due to its longitudinal nature, it is possible to examine trends and patterns at different times, giving researchers detailed information on how CVD develops and advances.

Dataset 2: The UCI ML repository provided the dataset, assembled from four different databases- Hungary, Cleveland, Switzerland, and VA Long Beach, providing a diverse range of samples making it more appropriate for analysis. This dataset was used for CVD prediction and exploring ML and ensemble learning models with 303 instances representing each by 14 attributes. Among these 14 attributes, one represents the target variable, and the rest 13 serve as features [31]. Moreover, the 7 attributes are continuous variables, denoting measurements or numerical data, while the remaining 6 are categorical variables, having discrete and limited possible values [32], [33]. In-depth analysis is possible with the integration of categorical and continuous variables, and the creation of efficient prediction models may help with the early detection and prevention of CVD.

Dataset 3: The 303 medical records making up the Z-Alizadeh Saini dataset provide important insights into cardiovascular health. Each record's 54 characteristics are divided into four groups: demographic data, symptoms and findings of a physical examination, ECG readings, blood test results, and the results of an echocardiogram [34]. The samples in the dataset are split into two classes: the normal class and the CAD (coronary artery disease) class [35]. Over half of the 303 samples—216 examples—belong to the CAD class, while the other 87 instances (28.71%) represent the regular class. A thorough understanding of the many causes and symptoms of coronary artery disease is provided by the combination of the 34 category aspects and 19 continuous features [36]. Researchers and medical practitioners may use this dataset to investigate relationships, identify trends, and create sophisticated prediction models that can help in the early identification and treatment of cardiovascular problems [37].

B. DATA PREPROCESSING

Part of the collected data may be obtained through the Internet, questionnaires, experiments, etc. It is common to encounter missing values in the data, which may be due to faulty instruments or human error. These missing and duplicate values can greatly reduce the number of training samples available for efficient model training, ultimately impacting the model's accuracy [38]. To this issue, mechanisms such as removing missing values from features and normalization

techniques have been employed to preprocess the datasets. Before the classification task can be applied, datasets have been normalized using the Min-Max scaler normalization approach that allows us to scale data in a dataset to a given range by utilizing each feature's minimum and maximum value.

C. HANDLE THE ISSUE OF OVERFITTING

A serious problem of overfitting occurs in Machine or Ensemble Learning models if trained with a small sample size. However, we have addressed this issue in a variety of ways.

1. The research used three benchmark datasets on CVD; to test the efficacy and accuracy of the proposed models. The data pre-processing techniques: removing null values, duplicate values, and min-max normalization techniques to convert the training and test data to a normal scale.

2. Four statistical FS techniques- Chi-Square, Gini Index, Information Gain, and ANOVA F-test, were used to select the optimal features and also significantly enhance the performance of the ensemble stacked learning models: stacking using SVM (SFS-SVM) and stacking using Cross-Validation Classifier (SFS-SCVC). These FS techniques address the curse of dimensionality and eliminate unimportant features.

3. Further, the stacked learning model using a Cross-Validation Classifier (SFS-SCVC) reduces training time and minimizes the risk of overfitting.

The proposed model has been tested rigorously on three benchmark datasets and has significantly improved the performance, and generalizability and enhanced the efficacy in a great way.

D. ML MODELS

1) LOGISTIC REGRESSION

It is a predictive statistical technique based on the probability idea used to classify dichotomous target variables as defined in (1). The name ''Logistic'' comes from the Logit function, the value of which lies between 0 and 1. The 'S-shaped' curve indicates the measure of the likelihood of whether the individuals have CVD [39].

$$f(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

where z can be modeled as a linear combination of features (x_i) with their corresponding weights (w_i) as illustrated below.

$$z = x_0 + x_1 w_1 + \ldots + x_k w_k \tag{2}$$

2) DECISION TREE

Unveiling the captivating realm of decision-making, the Decision Tree emerges as a compelling and insightful ML algorithm. Like the branches of a majestic tree reaching toward the sky, it navigates through complex datasets, unraveling patterns and uncovering hidden insights with unrivaled precision. Its strength lies in its ability to handle numerical and categorical data, gracefully adapting to diverse information landscapes. Each decision node, discerns the most informative features, guiding us toward enhanced understanding and informed choices [40]. The Decision Tree elegantly constructs a visual representation of knowledge, enabling us to traverse its branches and make well-informed predictions by intelligently partitioning the data based on key attributes [41]. Beyond its innate elegance, the Decision Tree is a versatile tool, capable of addressing challenges, from classification conundrums to regression riddles.

3) RANDOM FOREST

Random forest, a highly acclaimed and technologically sophisticated supervised ensemble classification method, has gained immense popularity [42], [43]. Its strength is developing a substantial number of trees during the training phase, resulting in an intense forest of decision trees from subsets. During testing, each tree within the forest independently assigns a class variable to individual data points. Ingeniously, the final determination for a given test data is made through a democratic process, where each tree's prediction contributes to a majority vote. The class variable securing the most votes qualifies as the most correct prediction contributing to the overall prediction accuracy of the test data [44], [45]. This strategy exploits the forest's collaborative knowledge, capturing tree diversity to minimize biases and enhance classification accuracy.

4) K NEAREST NEIGHBOR (KNN)

KNN is a non-parametric, instance-based learning system capable of generating predictions by measuring proximity. This ML approach can sense data and make informed decisions. Contrary to other techniques, it has no constraints on dataset size. So, it is more versatile when working with large datasets. It dynamically minimizes the complexity of the input by considering the k nearest neighbors in the feature space and combining their information for the classification of new instances. KNN is a good choice for pattern recognition and predictive modeling [46], [47] owing to its inherent adaptability and simplicity.

5) SUPPORT VECTOR CLASSIFIER

It is one of the well-performing ML methods to recognize complex patterns and make informed decisions. It generates a hyperplane using support vectors that optimally segregate multiple classes giving accurate predictions and informed decisions [48]. It processes data by linear or nonlinear, proving its adaptability and ability to manage large volumes of data. Due to its tendency to maximize the distance between classes, the SVC performs well and is noise- and outliers-resistant [49]. It's a popular choice having core qualities of refinement and accuracy.

6) XGBOOST

XGBoost supports data comprehension and well-informed decision-making. For handling vast datasets, execution speed should be high making it an ideal choice. It removes the

limitations by handling datasets of any size. Researchers work with data on a flexible scale that overcomes the constraints imposed by other algorithms [50]. At its core, XGBoost embodies the essence of a decision tree-based ensemble learning framework, leveraging the potency of Gradient Descent as its underlying objective function. This unique combination imparts a remarkable level of flexibility, ensuring optimal utilization of computational power to yield the desired results.

E. FEATURE SELECTION

The role of FS is paramount in ML as it addresses the curse of dimensionality and eliminates unimportant features. The objective is to opt for the most relevant attributes to enhance the accuracy and efficiency of the model. This reduces training time and minimizes the risk of overfitting [51]. By determining the features on which the output class label depends the most, the best features for the dataset can be identified. Redundant or correlated variables can hinder model generalization and decrease classifier accuracy [52], [53], [54]. The proposed method in this study combines different SFS techniques discussed below to determine the optimal features. The scores or ranks produced by these techniques are considered, and the union of the most contributing features is selected for the classifier.

1) INFORMATION GAIN

Information Gain, also known as Mutual Information, is a measure of the dependence between variables. It quantifies the shared information between the input (set of features) and the target variable in a dataset. Mutual information reveals how much knowing one variable reduces uncertainty about the other, especially, how much information is gained about the target variable by knowing the particular feature [55]. The purpose is to figure out which features in a set of learning feature vectors are the most successful at discriminating the classes to be learned [56], [57]. It is calculated using entropy to assess the importance of a given attribute in the feature vectors as shown in Equation 3 where p_i represents the probability of an attribute being classified for a distinct class label.

$$Entropy\,(X) = -\sum_{i=1}^{c} p_i log_2 p_i \tag{3}$$

Mathematically, the information gain (IG) of a feature X for target variable Y can be computed using the following Equation 4.

$$IG\,(X/Y) = H\,(Y) - H(Y/X) \tag{4}$$

The information gain *IG(X/Y)*represents how much information a specific feature X carries about the target variable Y. It is computed by comparing the entropy of the target variable with the conditional entropy $H(Y/X)$. The higher the information gain, the more entropy is removed, indicating the significance of variable X in the feature vector. To determine the importance of features, information gain is computed for every feature in the dataset. The features with the highest information gain are ranked first, followed by the features with the second highest information gain, and so on. These rankings reflect the relevance and importance of each feature in predicting the target variable [58]. Features with higher scores are considered more valuable and informative for the given task or problem, while features with lower scores may have less impact on the target variable.

2) GINI INDEX

It is a measure of the probability of misclassifying a specific feature when selected randomly, also known as Gini impurity which is computed by subtracting the sum of the squared probabilities of each class from one. It operates on categorical target variables, considering success or failure. It helps in determining the relevance of an attribute in classification tasks. As shown in equation 5, *Gini(X)* determines the gini index value for the attribute X representing n different classes and based on scores, ranks all the attributes from the feature vector which helps in determining the relevant features for the classifier.

$$Gini\,(X) = 1 - \sum_{i=1}^{n} p_i^2 \tag{5}$$

3) CHI-SQUARE

It is widely used in statistics for categorical elements in a dataset. To evaluate the interdependence of the variables, it examines the difference between the predicted and actual numbers using the *chi2()*function that is in the Sci-kit-learn library. The formula for computing chi-square requires comparing the observed frequency (O) and expected frequency (E) as shown in equation 6. Higher chi-square values show a more significant correlation between a feature and the outcome, resulting in important selection criteria. While a high chi-square value portrays association, and a strong correlation between the observed and expected values corroborates independence [59]. It succeeds well in determining the liaison between categorical indicators and aiding in FS for CVD patients. The last step is to decide on an acceptable threshold for the number of features to be a subset, i.e., the required number of features with the most significant Chi test score.

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{6}$$

4) ANOVA F-TEST

It is utilized to understand the variation of a particular categorical variable with a continuous variable. The variables are correlated if there is a variance between the different values or entries of the independent variable affecting the outcome.

So, the ANOVA f-test tests these variances depending on the statistical mean values of the groups [60]. A variance of the continuous variable concerning each group or for each unique value of the categorical variable i.e., how the continuous variable varies is estimated [61] [62]. If they do not vary,

then the predictor will be proved to be not correlated with the outcome as illustrated in Fig. 2.

So, the ANOVA f-test is performed considering each continuous variable from the dataset and ranks the predictor based on the F-statistic score (F_{ss}), which is the ratio of variance between groups (V_b) and variance within groups (V_w) as illustrated in equation 7. The features with higher ranks can be considered the optimal reduced features subset.

$$F_{ss} = \frac{V_b}{V_w} = \frac{sum\ of\ squares\ between\ groups}{sum\ of\ squares\ with\ groups} \tag{7}$$

ANOVA analysis test translates the data into two different estimates of population variance i.e., variance between groups and variance within groups, mathematically shown in Equations 8 and 9.

$$V_b = \frac{\sum_{i=1}^{n} n_i(\overline{x}_i - \overline{x})}{n-1} \tag{8}$$

$$V_w = \frac{\sum_{i=1}^{n} (x_i - \overline{x}_i)}{N-n} \tag{9}$$

where $\overline{x}_i$ the individual sample mean, $\overline{x}$ overall group mean, n_i number of entries in each group, n reflects the number of groups, x_i individual entry values and N is the total number of entries across all the groups. Optimal feature set (X_U) of each dataset is obtained from Union operation on a reduced subset of features selected based on scores from different statistical tests after considering the top most contributing features and removing the correlated ones. The mathematical representation is shown below.

$$\begin{aligned} X_{IGN} &= X_i.....X_p \\ X_{GNI} &= X_j.....X_q \\ X_{CHI} &= X_k.....X_r \\ X_{ANV} &= X_l.....X_s \\ X_U &= X_{IGN}\, UX_{GNI}\, UX_{CHI}\, UX_{ANV} \end{aligned} \tag{10}$$

F. STACKING MODELS

1) HYBRID SFS WITH STACKING USING SVM AS META-MODEL (SFS-SVM)

This technique involves two key steps: select the optimal feature set by union operation to the reduced features set based on scores obtained from four SFS techniques and use the SVM as the meta-model and the predictions of other base models. This stacking approach enhances the prediction by combining the contributions from six base ML models with the meta-model. The workflow of stacking with SVM as a meta-model is depicted in Fig. 3 showing the importance of combining the knowledge obtained from the basic models to produce a final prediction surpassing the constraints of any one model. This stacking method presents an effective framework for handling challenging issues and enhancing the ML model's general performance.

Stacking with SVM as meta-model algorithm 1 will train multiple base models on the full training set to create an ensemble model. By gathering the base model's predictions for every instance, it builds a training set for the meta-model and learns base models repeatedly. The meta-model determines the decision, and the basic model's predictions are combined to generate the ensemble model. This approach enhances the classification performance by leveraging the strengths of multiple models in the ensemble.

Algorithm 1 Stacking with SVM as Meta Model

Input: Training data TD
Output: An ensemble stacking model M
1. **Step 1**: Train base models on the entire training set
2. **for** k ← 1 to K **do**
3. Train a model M_k from TD
4. **end for**
5. **Step 2**: Create a training set for the meta-model
6. **for** x_i belonging to TD **do**
7. Obtain a value $\{m_i, y_i\}$, where $m_i = \{m1(x_i), m2(x_i), \ldots, m_k(x_i)\}$
8. **end for**
9. **Step 3:** Learn a meta-model
10. Train a new model m' from the pool of $\{m_i, y_i\}$
11. **Step 4:** Return the ensemble classifier M
12. $M(x) = m'(m1(x), m2(x), \ldots, m_k(x))$

2) HYBRID SFS WITH STACKING USING SCVC AS META-MODEL (SFS-SCVC)

This technique involves two key steps that lay the foundation for powerful ensemble learning. The first step focuses on selecting the optimal feature set by performing a union operation on the reduced set of features. These features are based on scores obtained from four SFS techniques. They are carefully chosen to ensure the inclusion of the most relevant and informative attributes. Building upon this foundation, the StackingCV Classifier takes center stage, traditionally, in the standard stacking procedure, the first-level classifiers are trained on the same dataset used to prepare inputs for the second-level classifier, potentially leading to overfitting. However, the StackingCV Classifier takes a different route by leveraging the concept of cross-validation. The dataset is divided into k folds, and in a series of successive rounds, k-1 folds are utilized to train the first-level classifiers. In each round, these trained classifiers are applied to the remaining subset not used for model fitting, generating stacked predictions and serving as the input data for the second-level classifier. This novel approach improves prediction accuracy and provides a more reliable and robust framework for using the collective insights of multiple classifiers. The StackingCV Classifier's workflow is shown in Fig. 4, highlighting the model's ability to stacking process. Due to the StackingCV Classifier's contribution for using all the available data, ML models can deliver better results and offer more informed CVD predictions.

The StackingCV Classifier, shown in Algorithm 2, is a meta-model that prepares the training set for the meta-model

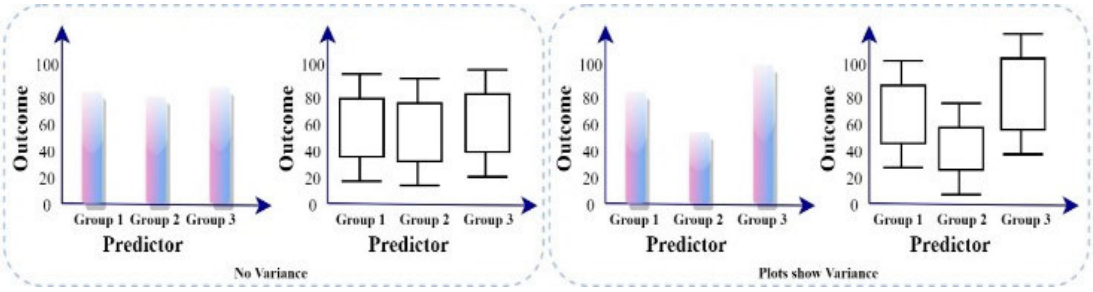

FIGURE 2. Bar Graphs and plot graphs showing predictor's impact on variance.

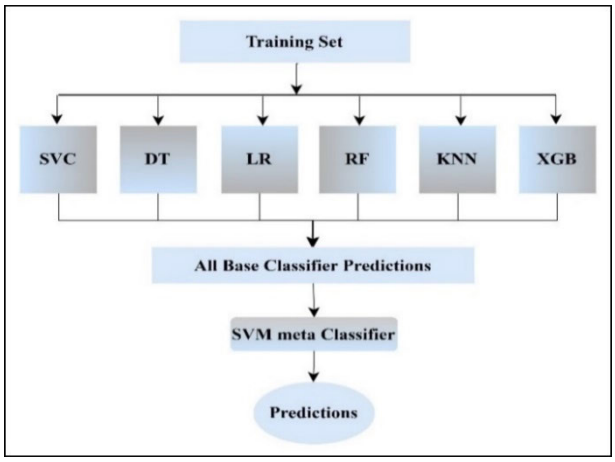

FIGURE 3. Stacking framework with SVM as meta model.

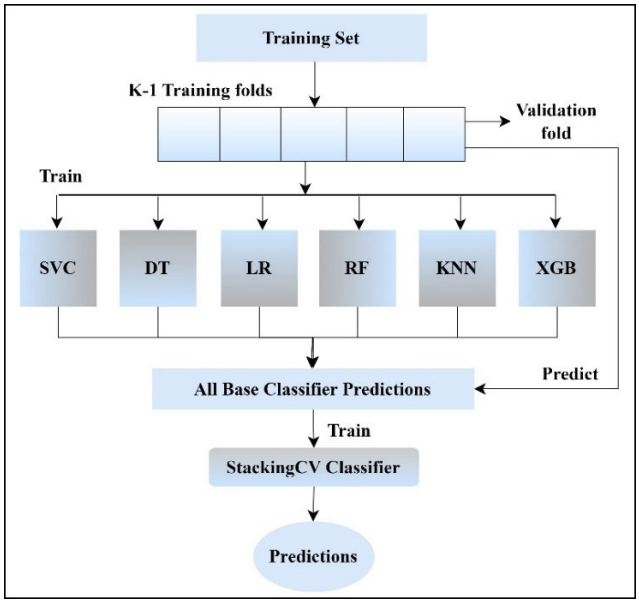

FIGURE 4. Stacking framework with StackingCV as meta model.

that uses a cross-validation. It uses the training data to train base models, gradually eliminating them one-fold at a time. The base model's predictions for every instance are then carefully compiled into a training set for the meta-model.

Algorithm 2 StackingCV Classifier as Meta Model

Input: Training data TD
Output: An ensemble stacking model N

Step 1: Apply CV in preparing the training set for the meta-model
Split TD into K subsets randomly: TD = {TD1, TD2, …, TD_k}
for p ← 1 to P **do**
Step 2: Learn base models
for q ← 1 to Q **do**
Train a model n_{pq} from TD \TD_k
end for
Step 3: Create a training set for the meta-model
for x_i belonging to TD_k **do**
Obtain a value {n_i, y_i}, where $n_i = \{n_{p1}(x_i), n_{p2}(x_i), \ldots, n_{pq}(x_i)\}$
end for
end for
Step 4: Train a meta-model
Train a new model n' from the pool of {n_i, y_i}
Step 5: Re-train base models
for i ← 1 to Q **do**
Train a model h_q based on the TD
end for
Step 6: Return the ensemble model N
N(x) = n'(n1(x), n2(x), …, n_q(x))

An ensemble classifier is created by combining the predictions from the base classifiers and using the meta-classifier to get the final answer. This unique method efficiently utilizes stacking and cross-validation to improve the performance and overall generalizability of the ensemble classifier.

IV. RESULTS AND DISCUSSION

A. IMPLEMENTATION DETAILS

Large dataset handling and effective model training are ensured by the hardware combination of an NVIDIA card (GeForce RTX 3070), an 11th Gen Intel Core i7-11700K processor, and 64GB of RAM. The 1.5TB disk capacity offers plenty of storage, while Ubuntu 22.04 LTS, a 64-bit operating system is the platform for resource management

and code execution. Jupyter Notebook was configured as the workstation.

The necessary Scikit-learn modules are imported for cross-validation, model stacking, and FS. The dataset is loaded into memory using Pandas, and data analysis techniques are applied to handle missing and deduped values. Seaborn and Matplotlib are used for data visualization, and NumPy is utilized for mathematical computations. FS techniques are applied to identify relevant features. Classification algorithms from Scikit-learn are used to implement ML models, and stacking models are built to improve predictive performance. Model evaluation is performed using appropriate metrics from the Scikit-learn library.

B. EVALUATION METRICS

Various evaluation metrics- accuracy, recall, precision, and f1-score have been put in place to measure the performance of the classification algorithm used in this research. All these measures shall be calculated based on the True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP) rates, using the confusing matrix set out in the confusion matrix table. TP, TN, FN, and FP were as expected by all models.

The disease is marked with the letters TP indicating that the patient has CVD, and the model also predicts this outcome, properly classifying a person with heart disease [63]. FN signals a patient's heart illness while the model predicted that the patient didn't have the illness; in other words, the model was misdiagnosed as not having it. FP is the outcome of the model incorrectly categorizing a healthy person as having cardiac disease when in fact the patient does not have the ailment. In the actual world, TN shows that the patient does not have cardiac disease, and the model agrees, correctly categorizing the patient as healthy and not anticipated to develop one in the future.

1) ACCURACY

It is the share of correctly categorized items that a trained ML model achieves, or the ratio of the number of correct predictions to the total number of predictions over every possible outcome, as can be shown by deriving a statistical measure for accuracy in Equation 11. A value between [0,100] and [1,0] is given depending on the scale used. Accuracy 0 indicates that the classifier routinely guesses the incorrect label, while accuracy 100 or 1 perpetually predicts the correct label [64], [65]. Values between these ranges depend on how well the classifier performs.

$$Accuracy\ (in\%) = \frac{TP + TN}{TP + TN + FP + FN} * 100 \tag{11}$$

2) PRECISION

It is one of the metrics of an ML model's performance that describes the proportion of patients successfully classified as having a CVD out of all the patients having it [66]. It is measured as the ratio of true positives to the total number of positive predictions, as shown mathematically in Equation 12.

$$Precision\ (in\%) = \frac{TP}{TP + FP} * 100 \tag{12}$$

3) RECALL

It is a metric to quantify the potential of a model to predict the positive instances. More positive samples become apparent when the recall value spikes. As a result, recall represents the proportion of individuals we accurately recognize as having heart disease out of all those who truly have the ailment [67], as shown mathematically in Equation 13.

$$Recall\ (in\%) = \frac{TP}{TP + FN} * 100 \tag{13}$$

4) F1-SCORE

For some cases, a high recall is more important than a high precision, such as detecting CVD patients. On the other hand, for events like classifying loan defaulters, a high precision is desired to avoid losing potential clients [42]. However, there are scenarios where both recall and precision are equally significant. The F1-score is the harmonic mean of precision and recall, is technically stated in Equation 14, and is used to assess the model's performance.

$$F1 - score\ (in\%) = 2 * \frac{Precision * Recall}{Precision + Recall} * 100 \tag{14}$$

C. PERFORMANCE EVALUATION

The performance of ML models (Logistic Regression, Decision Tree, Random Forest, KNN, SVM, and XG Boost.) without FS is evaluated using the three different datasets. The accuracy, precision, recall, and F1-score are reported against each model and dataset in Table 3.

The SVM model exhibits the highest accuracy values, ranging from 86.01% to 91.21%, with precision values varying between 88.47% and 93.62%. The comparative analysis of which is depicted in Fig. 5.

In Table 4, we explore the performance of the same ML models but with an optimal feature set. The results indicate that employing an optimal feature set improves the performance of ML models in terms of the metrics considered.

The SVM model consistently performs well across all datasets. Random Forest and XG Boost models also demonstrate competitive performances, the comparative analysis of which is represented in Fig. 6. The specific choice of the ML model and FS technique can greatly impact the predictive capabilities in different datasets, highlighting the importance of selecting appropriate features for accurate predictions.

The performance of Stacking models without FS is shown in Table 5. The first Stacking model employs SVM as the meta-model, while the second model utilizes SCVC as the meta-model. Both models are tested on three datasets. For the Stacking model with SVM as the meta-model, impressive results are observed, and the Stacking model with SCVC as the meta-model gives better results than the former for all

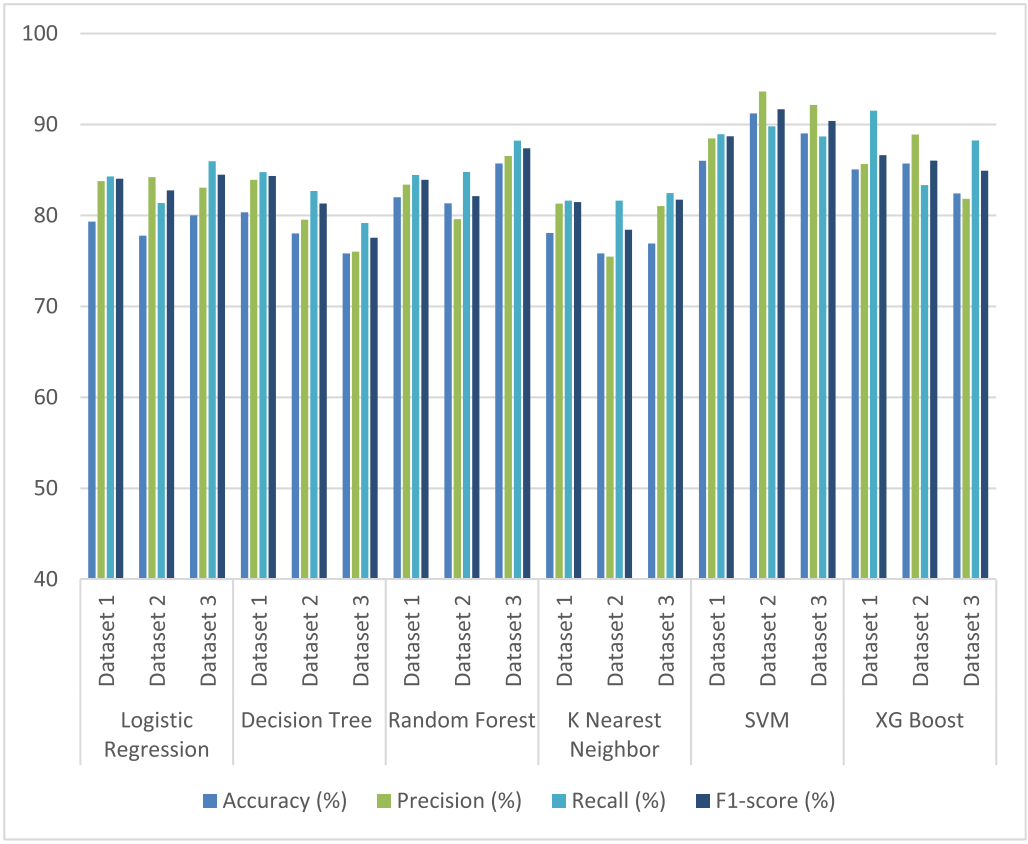

FIGURE 5. Comparative analysis of ML models without FS.

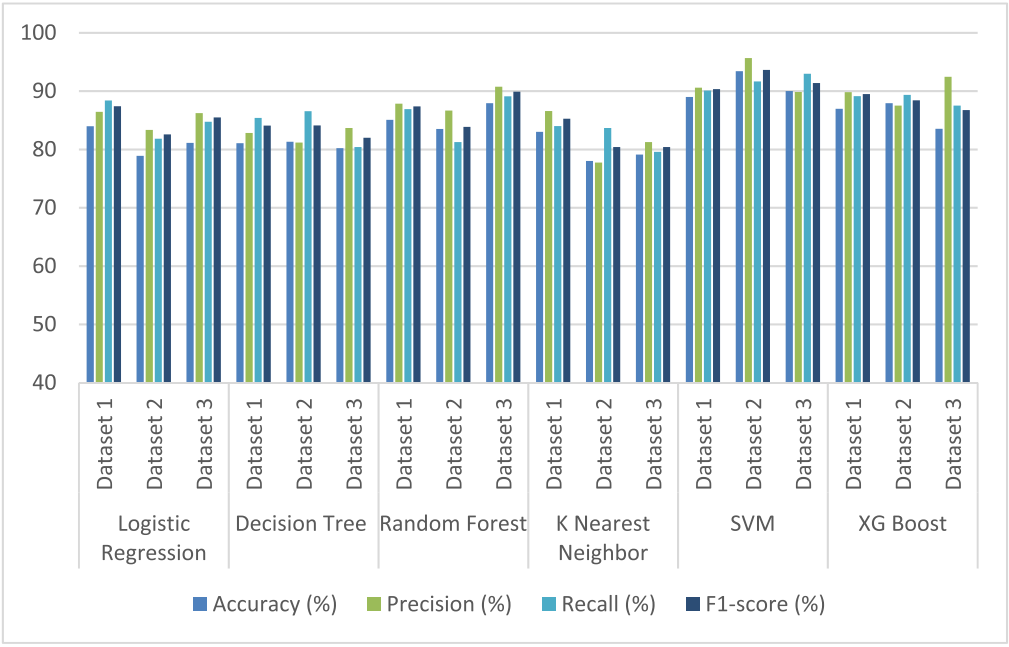

FIGURE 6. Comparative analysis of ML models with optimal feature set.

three datasets. The comparative analysis of the two stacking models with raw features is shown in Fig. 7.

The two proposed hybrid ensemble Stacked models with an optimal feature set are evaluated in Table 6. Both the proposed models are tested on the same three benchmark datasets. Approach 1 (SFS-SVM) demonstrates competitive performance across all datasets, achieving accuracy rates above 95% and precision, recall, and F1 scores

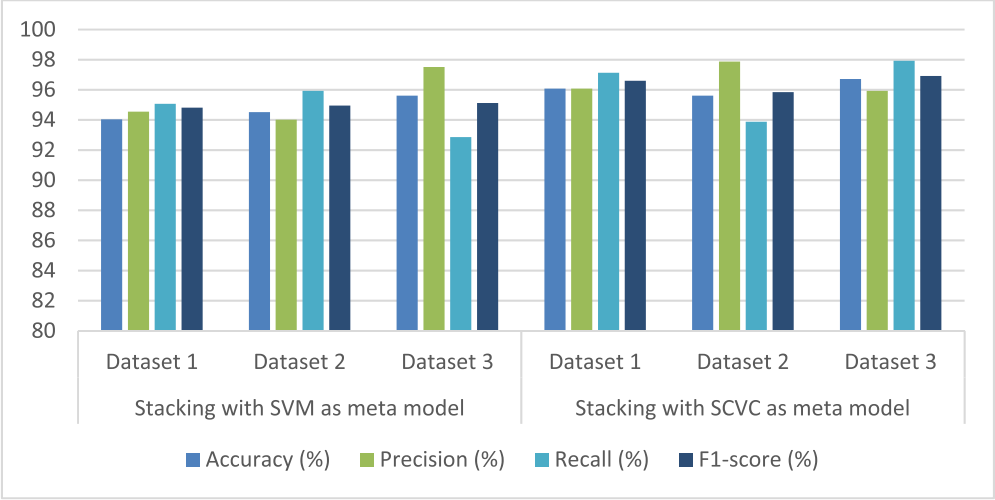

FIGURE 7. Comparative analysis of two stacking models without FS.

TABLE 3. Performance of ML models without feature selection in Percentage.

Models	Dataset	Accuracy	Precision	Recall	F1-score
Logistic Regression	Dataset 1	79.33	83.78	84.29	84.03
	Dataset 2	77.78	84.22	81.36	82.76
	Dataset 3	80.01	83.05	85.96	84.48
Decision Tree	Dataset 1	80.35	83.92	84.76	84.34
	Dataset 2	78.02	79.53	82.69	81.31
	Dataset 3	75.82	76.01	79.17	77.55
Random Forest	Dataset 1	81.99	83.38	84.44	83.91
	Dataset 2	81.32	79.59	84.78	82.11
	Dataset 3	85.71	86.54	88.23	87.38
K Nearest Neighbor	Dataset 1	78.07	81.29	81.62	81.46
	Dataset 2	75.82	75.47	81.63	78.43
	Dataset 3	76.92	81.03	82.46	81.74
SVM	Dataset 1	86.01	88.47	88.92	88.7
	Dataset 2	91.21	93.62	89.79	91.67
	Dataset 3	89.01	92.15	88.68	90.38
XG Boost	Dataset 1	85.06	85.65	91.52	86.62
	Dataset 2	85.71	88.89	83.34	86.02
	Dataset 3	82.42	81.82	88.24	84.91

TABLE 4. Performance of ML models with an optimal feature set in Percentage.

Models	Dataset	Accuracy	Precision	Recall	F1-score
Logistic Regression	Dataset 1	83.96	86.45	88.39	87.41
	Dataset 2	78.89	83.34	81.82	82.57
	Dataset 3	81.11	86.21	84.75	85.47
Decision Tree	Dataset 1	81.05	82.83	85.39	84.09
	Dataset 2	81.32	81.18	86.54	84.11
	Dataset 3	80.22	83.67	80.39	82
Random Forest	Dataset 1	85.06	87.85	86.92	87.38
	Dataset 2	83.51	86.67	81.25	83.87
	Dataset 3	87.91	90.74	89.09	89.9
K Nearest Neighbor	Dataset 1	83.01	86.57	84	85.27
	Dataset 2	78.02	77.73	83.67	80.39
	Dataset 3	79.12	81.25	79.59	80.41
SVM	Dataset 1	88.99	90.58	90.08	90.33
	Dataset 2	93.41	95.65	91.66	93.62
	Dataset 3	90.01	89.83	92.98	91.38
XG Boost	Dataset 1	86.95	89.82	89.14	89.48
	Dataset 2	87.91	87.51	89.36	88.42
	Dataset 3	83.52	92.45	87.51	86.73

exceeding 95%. Approach 2 (SFS-SCVC) outperforms the other models, showcasing remarkable accuracy, precision, recall, and F1-score values. Dataset 3 excels with an accuracy rate of 98.91%, precision of 97.16%, recall of 98.03%, and F1-score of 97.59%. A comparative analysis is presented in Figure 8, showcasing the performance of the two proposed hybrid techniques with an optimal feature set.

The proposed hybrid ensemble Stacked models with an optimal feature set Approach 2 (SFS-SCVC), combining SCVC with the optimal feature set obtained through SFS demonstrates exceptional accuracy, precision, recall, and F1-scores. SCVC utilizes cross-validation techniques to train and test the ensemble of base classifiers, ensuring robustness and generalization. This helps to mitigate overfitting and improves the model's capacity to produce precise forecasts on unobserved data.

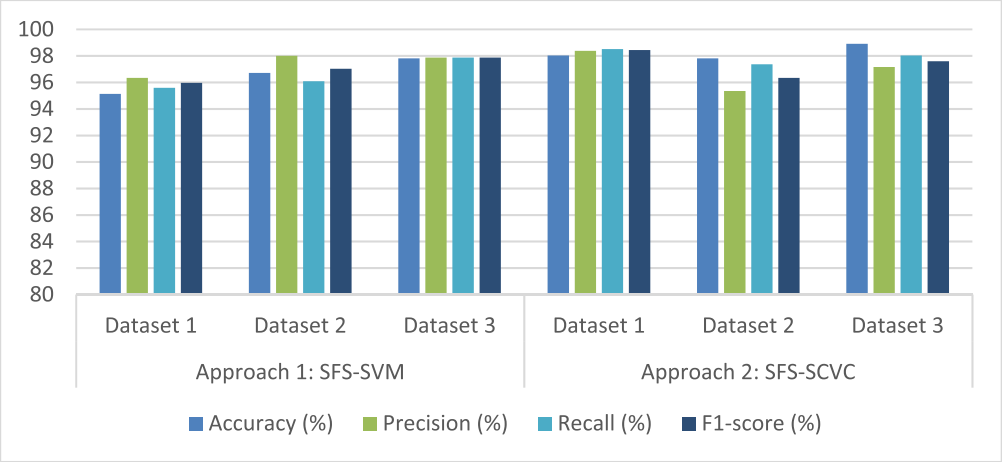

FIGURE 8. Comparative analysis of two Proposed Hybrid Ensemble Stacked models with an optimal feature set.

TABLE 5. Performance of two Stacking models without feature selection in Percentage.

Models	Dataset	Accuracy	Precision	Recall	F1-score
Stacking with SVM as meta model	Dataset 1	94.03	94.54	95.06	94.81
	Dataset 2	94.51	94.01	95.92	94.95
	Dataset 3	95.61	97.51	92.85	95.12
Stacking with SCVC as meta model	Dataset 1	96.07	96.07	97.13	96.59
	Dataset 2	95.61	97.87	93.88	95.84
	Dataset 3	96.71	95.92	97.92	96.91

TABLE 6. Performance of two Proposed Hybrid Ensemble Stacked models with an optimal feature set.

Models	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Approach 1: SFS-SVM	Dataset 1	95.13	96.34	95.59	95.97
	Dataset 2	96.71	98.01	96.08	97.03
	Dataset 3	97.81	97.87	97.87	97.87
Approach 2: SFS-SCVC	Dataset 1	98.03	98.38	98.51	98.44
	Dataset 2	97.81	95.35	97.36	96.34
	Dataset 3	98.91	97.16	98.03	97.59

These findings highlight the effectiveness of the Stacking and hybrid ensemble techniques in enhancing the performance of classification models by incorporating FS techniques. It identifies the most significant and applicable attributes in the dataset. Hence lowering the dimensionality and complexity of the input data. The models can focus on the most discriminative attributes set by selecting the optimal feature set, t, resulting in improved performance. In recent years different studies have been employed on CVD prediction using ML techniques. Table 7, provides a summary of recent studies conducted and compared with the proposed approach, along with their achieved results.

TABLE 7. Comparative analysis of existing and proposed methods.

Methods/Author(s)	Results (in Percentage)
Zhang et al. [18]	Accuracy: 98.56 Recall: 99.35 Precision: 97.84 F1-score: 98.3 AUC: 98.3
Jothi Prakash and Karthikeyan [20]	Accuracy: 93.65
Książek et al. [22]	Accuracy: 88.49 F1-score: 87.62
Venkatesan et al. [23]	Accuracy: 86.4
Robinson Spencer et al. [24]	Accuracy: 85
Atta Ur Rahman et al. [27]	Accuracy: 96.51
Proposed Approach 1 (SFS-SVM)	**Accuracy: 97.81** **Precision: 97.87** **Recall: 97.87** **F1-score: 97.87**
Proposed Approach 2 (SFS-SCVC)	**Accuracy: 98.91** **Precision: 97.16** **Recall: 98.03** **F1-score: 97.59**

V. CONCLUSION

In this paper, two compelling hybrid methods SFS-SVM and SFS-SCVC are proposed, that smoothly merge SFS concepts with the power of ML. These methods work as preventive tools for the early detection and diagnosis of CVD. The suggested work is divided into three key phases: a critical task of thorough dataset preparation dealing with missing and deduped values and using the min-max scaler for normalization. The second step focuses on the technique of FS, where the ideal feature set is methodically determined by combining feature sets produced from four different statistical methods: Chi-Square, Gini Index, Information Gain, and ANOVA F-test. To determine enhanced prediction skills, the effectiveness of two hybrid techniques that combine the use of optimum and raw feature sets with six well-known

ML models was evaluated in the final phase. Three benchmark datasets are included in the validation process: the Z-Alizadeh Saini dataset, the UCI ML Repository heart disease dataset, and the Framingham dataset. This allows for a thorough comparison study. The results showed that the two suggested hybrid ensemble stacked learning models, SFS-SVM and SFS-SCVC outperform traditional ML models in terms of accuracy while using fewer features. The SFS-SCVC model beat the SFS-SVM model with an accuracy of 98.03%, 97.81%, and 98.91% respectively on the three respective datasets. These techniques build trust in their dependability and demonstrate adaptability in accurately predicting CVD across various databases.

The proposed methods emerged as an innovative and trustworthy solution, amalgamating cutting-edge preprocessing techniques, FS prowess, and ingenious ensemble methodology. With its better performance and adaptability, it shows a significant advancement in the early diagnosis of CVD; thereby holding immense potential for real-world applications.

A. DATA AVAILABILITY

The data underlying this article are available at

- https://www.kaggle.com/datasets/aasheesh200/framingham -heart-study-dataset,
- https://archive.ics.uci.edu/dataset/45/heart +disease [DOI:10.24432/C52P4X],
- https://data.mendeley.com/datasets/vrymwyh2tg/1 [DOI: 10.17632/vrymwyh2tg.1]

B. CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

C. FUNDING STATEMENT

The authors received no specific funding for this study.

REFERENCES

[1] J. Stewart, G. Manmathan, and P. Wilkinson, "Primary prevention of cardiovascular disease: A review of contemporary guidance and literature," *JRSM Cardiovascular Disease*, vol. 6, Jan. 2017, Art. no. 204800401668721, doi: 10.1177/2048004016687211.

[2] R. Alizadehsani, M. J. Hosseini, A. Khosravi, F. Khozeimeh, M. Roshanzamir, N. Sarrafzadegan, and S. Nahavandi, "Non-invasive detection of coronary artery disease in high-risk patients based on the stenosis prediction of separate coronary arteries," *Comput. Methods Programs Biomed.*, vol. 162, pp. 119–127, Aug. 2018, doi: 10.1016/j.cmpb.2018.05.009.

[3] S. Mukhopadhay, A. Mukherjee, D. Khanra, B. Samanta, A. Karak, and S. Guha, "Cardiovascular disease risk factors among undergraduate medical students in a tertiary care centre of eastern india: A pilot study," *Egyptian Heart J.*, vol. 73, no. 1, p. 94, Oct. 2021, doi: 10.1186/s43044-021-00219-9.

[4] Y. Ruan, Y. Guo, Y. Zheng, Z. Huang, S. Sun, P. Kowal, Y. Shi, and F. Wu, "Cardiovascular disease (CVD) and associated risk factors among older adults in six low-and middle-income countries: Results from SAGE wave 1," *BMC Public Health*, vol. 18, no. 1, p. 778, Jun. 2018, doi: 10.1186/s12889-018-5653-9.

[5] P. Ghosh, S. Azam, M. Jonkman, A. Karim, F. M. J. M. Shamrat, E. Ignatious, S. Shultana, A. R. Beeravolu, and F. De Boer, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[6] Ş. Ay, E. Ekinci, and Z. Garip, "A comparative analysis of meta-heuristic optimization algorithms for feature selection on ML-based classification of heart-related diseases," *J. Supercomput.*, vol. 79, no. 11, pp. 11797–11826, Mar. 2023, doi: 10.1007/s11227-023-05132-3.

[7] A. Mahajan and B. Kaushik, "A review of machine learning algorithms and feature selection techniques for cardiovascular disease prediction: Insights and implications," in *Proc. 7th Int. Conf. Comput., Commun., Control Autom. (ICCUBEA)*, Aug. 2023, pp. 1–5, doi: 10.1109/iccubea58933.2023.10392135.

[8] M. Diwakar, A. Tripathi, K. Joshi, M. Memoria, P. Singh, and N. Kumar, "Latest trends on heart disease prediction using machine learning and image fusion," *Mater. Today, Proc.*, vol. 37, pp. 3213–3218, Jan. 2021, doi: 10.1016/j.matpr.2020.09.078.

[9] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," *Healthcare Anal.*, vol. 2, Nov. 2022, Art. no. 100060, doi: 10.1016/j.health.2022.100060.

[10] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network-genetic algorithm," *Comput. Methods Programs Biomed.*, vol. 141, pp. 19–26, Apr. 2017, doi: 10.1016/j.cmpb.2017.01.004.

[11] S. Akdağ, F. Kuncan, and Y. Kaya, "A new approach for congestive heart failure and arrhythmia classification using downsampling local binary patterns with LSTM," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 30, no. 6, pp. 2145–2164, Sep. 2022, doi: 10.55730/1300-0632.3930.

[12] E. Maini, B. Venkateswarlu, B. Maini, and D. Marwaha, "Machine learning–based heart disease prediction system for Indian population: An exploratory study done in south India," *Med. J. Armed Forces India*, vol. 77, no. 3, pp. 302–311, Jul. 2021, doi: 10.1016/j.mjafi.2020.10.013.

[13] R. Alizadehsani, M. Abdar, M. Roshanzamir, A. Khosravi, P. M. Kebria, F. Khozeimeh, S. Nahavandi, N. Sarrafzadegan, and U. R. Acharya, "Machine learning-based coronary artery disease diagnosis: A comprehensive review," *Comput. Biol. Med.*, vol. 111, Aug. 2019, Art. no. 103346, doi: 10.1016/j.compbiomed.2019.103346.

[14] V. Chaurasia and A. Chaurasia, "Novel method of characterization of heart disease prediction using sequential feature selection-based ensemble technique," *Biomed. Mater. Devices*, vol. 1, no. 2, pp. 932–941, Jan. 2023, doi: 10.1007/s44174-022-00060-x.

[15] T. Bikku, S. R. Nandam, and A. R. Akepogu, "A contemporary feature selection and classification framework for imbalanced biomedical datasets," *Egyptian Informat. J.*, vol. 19, no. 3, pp. 191–198, Nov. 2018, doi: 10.1016/j.eij.2018.03.003.

[16] K. Dissanayake and M. G. Md Johar, "Comparative study on heart disease prediction using feature selection techniques on classification algorithms," *Appl. Comput. Intell. Soft Comput.*, vol. 2021, pp. 1–17, Nov. 2021, doi: 10.1155/2021/5581806.

[17] D. Deepika and N. Balaji, "Effective heart disease prediction with grey-wolf with firefly algorithm-differential evolution (GF-DE) for feature selection and weighted ANN classification," *Comput. Methods Biomechanics Biomed. Eng.*, vol. 25, no. 12, pp. 1409–1427, Sep. 2022, doi: 10.1080/10255842.2022.2078966.

[18] D. Zhang, Y. Chen, Y. Chen, S. Ye, W. Cai, J. Jiang, Y. Xu, G. Zheng, and M. Chen, "Heart disease prediction based on the embedded feature selection method and deep neural network," *J. Healthcare Eng.*, vol. 2021, pp. 1–9, Sep. 2021, doi: 10.1155/2021/6260022.

[19] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Informat. Med. Unlocked*, vol. 19, 2020, Art. no. 100330, doi: 10.1016/j.imu.2020.100330.

[20] V. Jothi Prakash and N. K. Karthikeyan, "Enhanced evolutionary feature selection and ensemble method for cardiovascular disease prediction," *Interdiscipl. Sci., Comput. Life Sci.*, vol. 13, no. 3, pp. 389–412, Sep. 2021, doi: 10.1007/s12539-021-00430-x.

[21] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informat. Med. Unlocked*, vol. 16, Jan. 2019, Art. no. 100203, doi: 10.1016/j.imu.2019.100203.

[22] W. Ksiąüek, M. Abdar, U. R. Acharya, and P. Pławiak, "A novel machine learning approach for early detection of hepatocellular carcinoma patients," *Cognit. Syst. Res.*, vol. 54, pp. 116–127, May 2019, doi: 10.1016/j.cogsys.2018.12.001.

[23] M. Venkatesan, P. Lakshmipathy, V. Vijayan, and R. Sundar, "Cardiac disease diagnosis using feature extraction and machine learning based classification with Internet of Things(IoT)," *Concurrency Comput., Pract. Exper.*, vol. 34, no. 4, p. e6622, Feb. 2022, doi: 10.1002/cpe.6622.

[24] R. Spencer, F. Thabtah, N. Abdelhamid, and M. Thompson, "Exploring feature selection and classification methods for predicting heart disease," *Digit. Health*, vol. 6, Jan. 2020, Art. no. 205520762091477, doi: 10.1177/2055207620914777.

[25] Y. Kaya, F. Kuncan, and R. Tekin, "A new approach for congestive heart failure and arrhythmia classification using angle transformation with LSTM," *Arabian J. Sci. Eng.*, vol. 47, no. 8, pp. 10497–10513, Aug. 2022, doi: 10.1007/s13369-022-06617-8.

[26] N. Demir, M. Kuncan, Y. Kaya, and F. Kuncan, "Multi-layer co-occurrence matrices for person identification from ECG signals," *Traitement Signal*, vol. 39, no. 2, pp. 431–440, Apr. 2022, doi: 10.18280/ts.390204.

[27] A. U. Rahman, Y. Alsenani, A. Zafar, K. Ullah, K. Rabie, and T. Shongwe, "Enhancing heart disease prediction using a self-attention-based transformer model," *Sci. Rep.*, vol. 14, no. 1, p. 514, Jan. 2024, doi: 10.1038/s41598-024-51184-7.

[28] S. Balasubramaniam, C. V. Joe, C. Manthiramoorthy, and K. S. Kumar, "ReliefF based feature selection and gradient squirrel search algorithm enabled deep maxout network for detection of heart disease," *Biomed. Signal Process. Control*, vol. 87, Jan. 2024, Art. no. 105446, doi: 10.1016/j.bspc.2023.105446.

[29] A. Gupta, V. Jain, and A. Singh, "Stacking ensemble-based intelligent machine learning model for predicting post-COVID-19 complications," *New Gener. Comput.*, vol. 40, no. 4, pp. 987–1007, Dec. 2022, doi: 10.1007/s00354-021-00144-0.

[30] S. S. Mahmood, D. Levy, R. S. Vasan, and T. J. Wang, "The Framingham heart study and the epidemiology of cardiovascular disease: A historical perspective," *Lancet*, vol. 383, no. 9921, pp. 999–1008, Mar. 2014, doi: 10.1016/s0140-6736(13)61752-3.

[31] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–11, Jul. 2021, doi: 10.1155/2021/8387680.

[32] A. Tiwari, A. Chugh, and A. Sharma, "Ensemble framework for cardiovascular disease prediction," *Comput. Biol. Med.*, vol. 146, Jul. 2022, Art. no. 105624, doi: 10.1016/j.compbiomed.2022.105624.

[33] A. Bhowmick, K. D. Mahato, C. Azad, and U. Kumar, "Heart disease prediction using different machine learning algorithms," in *Proc. IEEE World Conf. Appl. Intell. Comput. (AIC)*, Jun. 2022, pp. 60–65, doi: 10.1109/AIC55036.2022.9848885.

[34] R. Valarmathi and T. Sheela, "Heart disease prediction using hyper parameter optimization (HPO) tuning," *Biomed. Signal Process. Control*, vol. 70, Sep. 2021, Art. no. 103033, doi: 10.1016/j.bspc.2021.103033.

[35] R. Alizadehsani, J. Habibi, M. J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, and Z. A. Sani, "A data mining approach for diagnosis of coronary artery disease," *Comput. Methods Programs Biomed.*, vol. 111, no. 1, pp. 52–61, Jul. 2013, doi: 10.1016/j.cmpb.2013.03.004.

[36] R. Alizadehsani, M. J. Hosseini, A. Khosravi, F. Khozeimeh, M. Roshanzamir, N. Sarrafzadegan, and S. Nahavandi, "Non-invasive detection of coronary artery disease in high-risk patients based on the stenosis prediction of separate coronary arteries," *Comput. Methods Programs Biomed.*, vol. 162, pp. 119–127, Aug. 2018, doi: 10.1016/j.cmpb.2018.05.009.

[37] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, "C-CADZ: Computational intelligence system for coronary artery disease detection using Z-Alizadeh sani dataset," *Appl. Intell.*, vol. 52, no. 3, pp. 2436–2464, Feb. 2022, doi: 10.1007/s10489-021-02467-3.

[38] R. Alizadehsani, A. Khosravi, M. Roshanzamir, M. Abdar, N. Sarrafzadegan, D. Shafie, F. Khozeimeh, A. Shoeibi, S. Nahavandi, M. Panahiazar, A. Bishara, R. E. Beygui, R. Puri, S. Kapadia, R.-S. Tan, and U. R. Acharya, "Coronary artery disease detection using artificial intelligence techniques: A survey of trends, geographical differences and diagnostic features 1991–2020," *Comput. Biol. Med.*, vol. 128, Jan. 2021, Art. no. 104095, doi: 10.1016/j.compbiomed.2020.104095.

[39] C. Pan, A. Poddar, R. Mukherjee, and A. K. Ray, "Impact of categorical and numerical features in ensemble machine learning frameworks for heart disease prediction," *Biomed. Signal Process. Control*, vol. 76, Jul. 2022, Art. no. 103666, doi: 10.1016/j.bspc.2022.103666.

[40] Y. F. Khan, B. Kaushik, C. L. Chowdhary, and G. Srivastava, "Ensemble model for diagnostic classification of Alzheimer's disease based on brain anatomical magnetic resonance imaging," *Diagnostics*, vol. 12, no. 12, p. 3193, Dec. 2022, doi: 10.3390/diagnostics12123193.

[41] V. Chang, V. R. Bhavani, A. Q. Xu, and M. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," *Healthcare Anal.*, vol. 2, Nov. 2022, Art. no. 100016, doi: 10.1016/j.health.2022.100016.

[42] P. R. L, S. V. Jinny, and Y. V. Mate, "Early prediction model for coronary heart disease using genetic algorithms, hyper-parameter optimization and machine learning techniques," *Health Technol.*, vol. 11, no. 1, pp. 63–73, Jan. 2021, doi: 10.1007/s12553-020-00508-4.

[43] A. Chunduru, A. R. Kishore, B. K. Sasapu, and K. Seepana, "Multi chronic disease prediction system using CNN and random forest," *Social Netw. Comput. Sci.*, vol. 5, no. 1, p. 157, Jan. 2024, doi: 10.1007/s42979-023-02521-6.

[44] S. Kapoor, L. Kasar, A. Mandole, and J. Mahajan, "Heart disease prediction using machine learning and data analytics approach," in *Proc. 7th Int. Conf. Comput. Eng. Technol. (ICCET)*, vol. 2022, Feb. 2022, pp. 357–360, doi: 10.1049/icp.2022.0647.

[45] Y. F. Khan and B. Kaushik, "Neuro-image classification for the prediction of Alzheimer's disease using machine learning techniques," in *Proc. Int. Conf. Mach. Intell. Data Sci. Appl.*, M. Prateek, T. P. Singh, T. Choudhury, H. M. Pandey, and N. G. Nhu, Eds. Singapore: Springer, 2021, pp. 483–493, doi: 10.1007/978-981-33-4087-9_41.

[46] N. M. Idris, Y. K. Chiam, K. D. Varathan, W. A. W. Ahmad, K. H. Chee, and Y. M. Liew, "Feature selection and risk prediction for patients with coronary artery disease using data mining," *Med. Biol. Eng. Comput.*, vol. 58, no. 12, pp. 3123–3140, Dec. 2020, doi: 10.1007/s11517-020-02268-9.

[47] M. Mijwil and B. Salman Shukur, "A scoping review of machine learning techniques and their utilisation in predicting heart diseases," *Ibn AL-Haitham J. Pure Appl. Sci.*, vol. 35, no. 3, pp. 175–189, Jul. 2022, doi: 10.30526/35.3.2813.

[48] A. Abdellatif, H. Abdellatef, J. Kanesan, C.-O. Chow, J. H. Chuah, and H. M. Gheni, "An effective heart disease detection and severity level classification model using machine learning and hyperparameter optimization methods," *IEEE Access*, vol. 10, pp. 79974–79985, 2022, doi: 10.1109/ACCESS.2022.3191669.

[49] M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," *Healthcare Anal.*, vol. 3, Nov. 2023, Art. no. 100130, doi: 10.1016/j.health.2022.100130.

[50] Y. Zhang, R. Lin, H. Zhang, and Y. Peng, "Vibration prediction and analysis of strip rolling mill based on XGBoost and Bayesian optimization," *Complex Intell. Syst.*, vol. 9, no. 1, pp. 133–145, Feb. 2023, doi: 10.1007/s40747-022-00795-6.

[51] R. Alizadehsani, M. J. Hosseini, R. Boghrati, A. Ghandeharioun, F. Khozeimeh, and Z. A. Sani, "Exerting cost-sensitive and feature creation algorithms for coronary artery disease diagnosis," *Int. J. Knowl. Discovery Bioinf.*, vol. 3, no. 1, pp. 59–79, Jan. 2012, doi: 10.4018/jkdb.2012010104.

[52] R. Alizadehsani, M. Roshanzamir, M. Abdar, A. Beykikhoshk, A. Khosravi, S. Nahavandi, P. Plawiak, R. S. Tan, and U. R. Acharya, "Hybrid genetic-discretized algorithm to handle data uncertainty in diagnosing stenosis of coronary arteries," *Expert Syst.*, vol. 39, no. 7, Aug. 2022, Art. no. e12573, doi: 10.1111/exsy.12573.

[53] D. Panda, R. Ray, A. A. Abdullah, and S. R. Dash, "Predictive systems: Role of feature selection in prediction of heart disease," *J. Phys., Conf. Ser.*, vol. 1372, no. 1, Nov. 2019, Art. no. 012074, doi: 10.1088/1742-6596/1372/1/012074.

[54] S. Diwan, G. S. Thakur, S. K. Sahu, M. Sahu, and N. K. Swamy, "Predicting heart diseases through feature selection and ensemble classifiers," *J. Phys., Conf. Ser.*, vol. 2273, no. 1, May 2022, Art. no. 012027, doi: 10.1088/1742-6596/2273/1/012027.

[55] A. K. Verma, S. Pal, and B. B. Tiwari, "Skin disease prediction using ensemble methods and a new hybrid feature selection technique," *Iran J. Comput. Sci.*, vol. 3, no. 4, pp. 207–216, Dec. 2020, doi: 10.1007/s42044-020-00058-y.

[56] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[57] S. J. Sushma, T. A. Assegie, D. C. Vinutha, and S. Padmashree, "An improved feature selection approach for chronic heart disease detection," *Bull. Electr. Eng. Informat.*, vol. 10, no. 6, pp. 3501–3506, Dec. 2021, doi: 10.11591/eei.v10i6.3001.

[58] C. B. Gokulnath and S. P. Shantharajah, "An optimized feature selection based on genetic approach and support vector machine for heart disease," *Cluster Comput.*, vol. 22, no. S6, pp. 14777–14787, Nov. 2019, doi: 10.1007/s10586-018-2416-4.
[59] R. R. Sarra, A. M. Dinar, M. A. Mohammed, and K. H. Abdulkareem, "Enhanced heart disease prediction based on machine learning and χ^2 statistical optimal feature selection model," *Designs*, vol. 6, no. 5, p. 87, Sep. 2022, doi: 10.3390/designs6050087.
[60] N. O. F. Elssied, O. Ibrahim, and A. H. Osman, "A novel feature selection based on one-way anova f-test for e-mail spam classification," *Res. J. Appl. Sci., Eng. Technol.*, vol. 7, no. 3, pp. 625–638, Jan. 2014, doi: 10.19026/rjaset.7.299.
[61] J. Hassannataj Joloudari, F. Azizi, M. A. Nematollahi, R. Alizadehsani, E. Hassannatajjeloudari, I. Nodehi, and A. Mosavi, "GSVMA: A genetic support vector machine ANOVA method for CAD diagnosis," *Frontiers Cardiovascular Med.*, vol. 8, pp. 1–15, Feb. 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcvm.2021.760178
[62] G. Tripathy and A. Sharaff, "AEGA: Enhanced feature selection based on ANOVA and extended genetic algorithm for online customer review analysis," *J. Supercomput.*, vol. 79, no. 12, pp. 13180–13209, Aug. 2023, doi: 10.1007/s11227-023-05179-2.
[63] U. Nagavelli, D. Samanta, and P. Chakraborty, "Machine learning technology-based heart disease detection models," *J. Healthcare Eng.*, vol. 2022, pp. 1–9, Feb. 2022, doi: 10.1155/2022/7351061.
[64] A. Mahajan and B. Kaushik, "A data preprocessing and stacking ensemble learning model for improved CHD prediction," in *Advances in Mathematical Modelling, Applied Analysis and Computation*, J. Singh, G. A. Anastassiou, D. Baleanu, and D. Kumar, Eds. Cham, Switzerland: Springer, 2024, pp. 249–258, doi: 10.1007/978-3-031-56304-1_16.
[65] B. Kaushik and H. Banka, "Performance evaluation of approximated artificial neural network (AANN) algorithm for reliability improvement," *Appl. Soft Comput.*, vol. 26, pp. 303–314, Jan. 2015, doi: 10.1016/j.asoc.2014.10.002.
[66] M. Ashok and A. Gupta, "A systematic review of the techniques for the automatic segmentation of organs-at-risk in thoracic computed tomography images," *Arch. Comput. Methods Eng.*, vol. 28, no. 4, pp. 3245–3267, Sep. 2020, doi: 10.1007/s11831-020-09497-z.
[67] A. Chadha and B. Kaushik, "A survey on prediction of suicidal ideation using machine and ensemble learning," *Comput. J.*, vol. 64, no. 11, pp. 1617–1632, Nov. 2019, doi: 10.1093/comjnl/bxz120.

ABHIGYA MAHAJAN received the B.E. degree in the discipline of computers from the University of Jammu, Jammu and Kashmir, India, and the M.Tech. degree in computer science from the Department of Computer Science and IT, University of Jammu. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Shri Mata Vaishno Devi University, Katra, Jammu and Kashmir. He has cleared the GATE and UGC-NET exams. He also holds a teaching experience of three years. His research interests include artificial intelligence, machine learning, and nature-inspired algorithms.

BAIJNATH KAUSHIK received the B.E. degree in computer science and engineering from Nagpur University, Nagpur, in 1997, the Master of Technology degree from the School of Information Technology, GGSIPU, New Delhi, in 2009, and the Ph.D. degree in computer science from IIT Dhanbad, Dhanbad, in 2016. He is currently an Associate Professor and the Head of the School of Computer Science and Engineering, Shri Mata Vaishno Devi University, Katra, Jammu and Kashmir. His research interests include machine learning, deep learning, nature-inspired algorithms, soft computing, and parallel algorithms. He has published more than 80 research papers in the reputed SCI/SCIE, Scopus, and Web of Science journals and conferences. He had notably published and received the certificates for the Indian and Foreign patents.

MOHAMMAD KHALID IMAM RAHMANI (Senior Member, IEEE) was born in Patherghatti, Kishanganj, Bihar, India, in 1975. He received the B.Sc. (Engg.) degree in computer engineering from Aligarh Muslim University, India, in 1998, the M.Tech. degree in computer engineering from Maharshi Dayanand University, Rohtak, in 2010, and the Ph.D. degree in computer science engineering from Mewar University, India, in 2015. From 1999 to 2006, he was a Lecturer with the Maulana Azad College of Engineering and Technology, Patna. From 2006 to 2008, he was a Lecturer and a Senior Lecturer with the Galgotias College of Engineering and Technology, Greater Noida. From 2010 to 2011, he was an Assistant Professor with MVN, Palwal. He is currently an Associate Professor with the Department of Computer Science, College of Computing and Informatics, Saudi Electronic University, Riyadh, Saudi Arabia. He has published more than 70 research papers in journals and conferences of international repute, three book chapters, and holds one USA patent and another Australian patent of innovation. His research interests include algorithms, the IoT, cryptography, image retrieval, pattern recognition, machine learning, and deep learning.

ABDULBASID S. BANGA has been an Assistant Professor with the College of Computing and Informatics, Saudi Electronic University, Riyadh, Saudi Arabia, since 2014. He has more than 18 years of academic experience teaching in various universities in India. Previously, he was an Assistant Professor with the GLS Institute of Computer Technology (MCA Course), Ahmedabad, India. He has vast experience teaching at national and international levels. He is associated with various technical societies of national and international reputation. He has published various research articles in reputable journals. He is a Life Member of the Computer Society of India (CSI), profoundly engrossed in software estimation models, software engineering, data science, machine learning, artificial intelligence, blockchain technology, the Internet of Things, and cloud computing.