IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 November 2023, accepted 27 November 2023, date of publication 30 November 2023,
date of current version 11 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3338369

RESEARCH ARTICLE

Forecasting Coronary Heart Disease Risk With a 2-Step Hybrid Ensemble Learning Method and Forward Feature Selection Algorithm

SUSHREE CHINMAYEE PATRA, B. UMA MAHESWARI, (Senior Member, IEEE), AND PEETA BASA PATI, (Senior Member, IEEE)
Department of Computer Science and Engineering, Amrita School of Computing, Bengaluru, Amrita Vishwa Vidyapeetham, Bengaluru, Karnataka 560035, India
Corresponding author: B. Uma Maheswari (b_uma@blr.amrita.edu)

ABSTRACT Detecting cardiovascular irregularities in a timely manner is crucial for preventing any fatal risks. This research aims to devise an efficient forecasting algorithm for the timely prognosis of Coronary Heart Disease (CHD). The study includes a diverse sample of individuals from Framingham, Massachusetts, with varying demographic, clinical, and co-morbidity parameters. We aim to achieve this with a two-step ensemble Machine Learning model. Firstly, feature importance is integrated with conventional classifiers to build Feature Weighted Meta-Models with a Forward feature selection algorithm. Subsequently, the top-performing Meta-Models are combined to design the Hybrid Voting Models to predict the risk of CHD in a ten-year timeframe by minimizing the misclassification rate. The proposed models undergo vetting using multiple metrics, including F1 score, Matthew's Correlation Coefficient (MCC), Misclassification Ratio (MCR), and Accuracy. Given the high cost associated with misclassification in the healthcare domain, these metrics are carefully considered. The resulting model demonstrated strong predictive capability for CHD risk, achieving an overall accuracy rate of 95.87%. The F1 score is calculated to be 0.91, the MCC is 0.83, and the MCR is 0.041. Notably, the model achieved these impressive results using only seven features, reducing the time complexity of the prediction. In comparison to conventional classifiers, our model achieved a 23.94% improvement in accuracy, and a 17.23% improvement over average Meta-models accuracy, highlighting its effectiveness in predicting CHD risk.

INDEX TERMS Accuracy, ensemble learning, feature weighted meta-models, F1 score, feature importance, feature optimization, hybrid voting model, machine learning, MCC, misclassification ratio, time complexity.

I. INTRODUCTION

Healthcare is one of the greatest fields in which many academics have shown true interest. In order to deliver effective and affordable healthcare services, there is currently a lot of focus on restructuring existing healthcare processes utilizing various technological techniques. Automation in healthcare attempts to provide doctors with a wealth of patient information in an efficient manner. For the scientific community, it opens up a completely new field of investigation, and several studies are being done in this area. Furthermore, the exposure of the global demographics to lifestyle diseases namely Hypertension, Diabetes, Risk of Cardiovascular disease, etc. is at an all-time high owing to various economic and professional factors leading to higher risk for the human heart [1]. Coronary Heart Disease (CHD) is a leading cause of morbidity and mortality worldwide [2]. According to statistics from WHO, nearly 18 million people succumb to various heart diseases every year globally contributing to around 32% of the overall global death. Accurate prediction of risk is crucial for the prevention and management of this condition. Traditional risk prediction methods, such as the Framingham Risk Score, have limitations in terms of their predictive ability and generalizability to diverse populations [3]. On the other

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

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

hand, Machine Learning techniques can potentially improve risk prediction by leveraging large datasets and a variety of risk factors.

In this study, we aim to evaluate the performance of various Machine Learning classifiers and ensemble learning models in predicting the risk of CHD. We used a dataset containing information on gender, age, education level, body mass index (BMI), smoking history, heart rate, glucose level, and history of heart disease, hypertension, and diabetes, as well as CHD outcomes in a time frame of 10 years. We obtained the relative importance of considered risk factors using the Random Forest feature importance technique and provided the input features iteratively to the ML classifiers based on their ranking of importance. The Machine Learning models included k-nearest neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Extra Tree (ET), AdaBoost (AdB), and XGBoost (XGB). Further, we trained a two-layer ensemble model to foster our prediction by minimizing the blind spot of constituent classifiers. The first layer consisted of a Feature weighted Meta-model designed with a Forward Feature Selection algorithm. For the second layer, we selected the top performing Meta-models to build our proposed Hybrid Voting Model to predict the risk of CHD. In doing so, we combined an appropriate feature selection algorithm with a multi-step ensemble predictor in order to strengthen the outcomes by reducing the overall rate of misclassification of CHD risk prediction over a 10-year time frame.

The findings of this study have the potential to inform the development of more accurate and efficient risk prediction tools for CHD, which can be used in clinical practice to identify individuals at high risk and guide preventive interventions. By doing so, we hope to contribute to the development of more accurate and reliable CHD risk prediction models that can aid in the prevention and management of this important public health issue. Our research also aligns with the 3rd Sustainability Development Goal (UNSDG-3) the United Nations rolled out, emphasizing the importance of a healthy lifestyle and well-being for humans of all ages and demography.

Our proposed research aims to devise a highly effective model to predict the risk of human cardiovascular disease. We aim to achieve this by creating a 2-step Hybrid Ensemble model that is built on the principle of shielding the blind spots of standalone classifiers and converging for a better predictive outcome. The above proposition also factors in our cognizance of the staggering cost of misclassification in the healthcare sector and strives to produce a more accurate result as measured by several metrics. We also take the cost of computational complexity into account and hence another objective of this work is to optimize the model for superior results with a smaller number of input features. Our key contributions to the research are listed below:

- The 2-step ensemble Machine Learning classifier involved building Feature Weighted Meta-Models and subsequently combining top-performing models to design the Hybrid Voting Model
- Efficient reduction of individual classifiers' blind spots to eliminate misclassification with the Hybrid Ensemble model
- Iterative incorporation of relative feature importance using a Forward Feature Selection Algorithm within a 2-step Ensemble classifier to optimize computational complexity

The organization of the rest of the paper is outlined below. Section II covers a comprehensive overview of the latest research in the field. In Section III, we delve into the various feature engineering methods to obtain the optimal feature set as well as the design of our proposed Feature weighted Meta-models and Hybrid Voting Model. The results and analysis of our experiments are presented in Section IV where we have described the inference drawn from the evaluation parameters of various models and presented a benchmarking study. Finally, in Section V, we summarize our findings and discuss potential future directions for this research.

II. LITERATURE SURVEY

Healthcare systems intend to leverage the rise in connectivity and mobile device infrastructures in employing smart devices [4] capable of establishing an IoT network. Smart devices are immensely useful for monitoring the patient's condition, analyzing medical information, and proactively detecting the degree of involvement expected from the doctor. In recent years, the number of healthcare applications using wearable technology has significantly increased [5] owing to the advantage of being comfortable over conventional medical devices. One such example is a smart device that consolidates ECG, Accelerometer, and SPO2 in a single form factor [6]. These devices use various sensors to collect data about the human body. Working with sensors has its own set of limitations with respect to their energy consumption and transmission loss of data packets, as the patient cannot be in idle condition for a long time. Hence, it is important to place an intermediate node in the path of transmission to enhance the efficacy of the network concerning energy utilization, cost, and latency [7], [8].

Without a doubt, the advancements that IoT has achieved in the healthcare industry are unfathomable [9]. It can aid in the early diagnosis of some asymptomatic conditions like anemia and breast cancer. A quick first-aid response time, an effective communication system, and a user-friendly interface are three key components that must be considered when developing effective healthcare equipment or systems for the senior population [10]. The adoption of IoT-based software largely powers the influx of big data and allied techniques in the healthcare sector. Fahad et al. [11] have discussed how this software is useful for the collection of large-scale biometric data to predict the health statistics of patients. Ravindran et al. [12] have devised a novel method by combining Apache Spark and deep learning in big data to predict the physical risk and probability of a patient getting readmitted to the hospital. Predictive analysis has gained further understanding with the use of diverse Machine Learning models for healthcare big data, as elaborated by Daliya et al. [13].

Researchers have indicated the usage of deep computation methods on ECG data for the proactive detection of Cardiovascular disease risk [14]. In [15] and [16] the authors have focused on predicting cardiovascular risk within a 10-year time frame in the presence of comorbidity conditions for Asian population data. Even real-time ECG data cannot perfectly recreate the complexity of the human heart because it is such a complicated organ. As a result, it's important to recognize and give the platform permission to collect various sorts of data from various sensors. Therefore, Farman et al. [17] have gone a step ahead to predict the risk associated with the human heart using ensemble computational methods on fused features gathered from sensors and the medical history of the patient. Khanna et al. [18] have fused conventional CHD risk factors with an ultrasound image of carotid to predict the irregularities associated with it; this further led to the research done by Jain et al. [19] that focuses on ML-enabled biomarkers to prevent CHD risk.

Contemporary research by Yar et al. [20] has centered on the importance of feature engineering of the input parameters that go into the deep computational models for better prediction accuracy of Cardio Vascular Disease (CVD). Researchers have also performed ANN modeling on molecular diagnostics data to predict intra-organ failure of CVD patients and subsequently have forecasted the recovery rate using various Machine Learning algorithms [21]. The authors in [22] have proposed an ingenious method to analyze cardiac MRI with the help of disease classification and segmentation.

According to Wongvibulsin et al. [23], studies conducted with patients in a connected environment to monitor various risk factors of CVD, e.g., dietary practices, lipid profile, smoking habit, and morbidity history, showed encouraging results when encountered in a computational model. Wanda-CVD [24] acted as a remote health monitoring system that worked on a smartphone and demonstrated great results in educating the women population about identified CVD risks through feedback on their lifestyle.

The foremost step in the design of an analytical model to foresee the risk of CHD is the selection of risk factors that will be used as input for the model. The authors in [25], and [26] have highlighted several risk factors in their research that make for the risk of heart disease in various stages of events of CHD. Giardina et al. [27] have focused on developing a supervised kNN model to predict the risk of CHD for the Type-2 Diabetic population with 64% effectiveness. The impact of hypertension on CHD has been captured in [28] as a part of their research to predict the risk of CHD in a 3-year time frame for hypertension patients.

Recent research done by Krishnani et al. [29] describes the importance and methods of feature engineering while working with large-scale healthcare data for ML classifiers. The experiments conducted by the authors in [30], [31], and [32] highlight the impact input features may have on the overall prediction model developed using Machine Learning techniques and test their models with various feature extraction techniques. A study undertaken by Nithya et al. [33] highlighted the importance of Machine Learning-based models to efficiently analyze healthcare data and focuses on indispensable movement in this direction to provide personalized healthcare. In their recent research, Hassan et al. [34] have implemented various Machine Learning-based classifiers to forecast the risk of CHD. Another analytical model designed by Bemando et al. [35] on public data from the Cleveland database using Random Forest and Naive Bayes (Gaussian and Bernoulli) on various risk factors compared the results across these models.

Table 1 describes state-of-the-art studies on the usage of Machine Learning-based classifiers to predict the risk of CHD.

TABLE 1. Snapshot of state-of-the-art research on CHD risk evaluation.

Author	Algorithms used	Dataset used	Accuracy
Hassan *et al.* [34], 2022	NN, LR, SVM, XGB, NB, RF, DT, RBF, GBT, KNN, MLP	Heart Dataset (UCI repository)	96.28%, Random Forest
Truong *et al.* [38], 2022	CART, LDA, AB, ET, LR, SVM, MNB, XGB, RF	Heart Disease Dataset (UCI repository)	90%, AdaBoost
Abdalrada *et al.* [39], 2022	NB, SVM, DT	Heart Disease Dataset (UCI repository)	90%, Decision Tree
Singh *et al.* [40], 2022	KNN, SVM, DT, NB, LR	Heart Disease Dataset (UCI repository)	92%, Linear Regression
Kishor *et al.* [44], 2020	LM, LR, NB, DT, RF, HRFLM SVM	Heart Cleveland (UCI repository)	88.40%, Hybrid Random Forest Linear Model
Bemando *et al.* [35], 2021	RF, Gaussian-NB, Bernoulli-NB	UCI-Cleveland database	85%, Naive Bayes
Gupta *et al.* [37], 2022	DT, RF, LR	UCI-Cleveland database	92.10%, Linear Regression
Arumugam *et al.* [36], 2021	LR, NB, M5P Tree, RF REP, J48, JRIP	Hungarian and Statlog (heart) dataset	99.81%, Random Forest
Doppala *et al.* [41], 2022	Random Forest into fetal echocardiography	Congenital heart disease database with 3910 Singleton Fetuses	Sensitivity 85% and Specificity 88%
Gudmundsson *et al.* [42], 2022	kNN,LR, SVM, RF	Pathogen, Host feature	99%, Random Forest
Marco *et al.* [44], 2018	kNN, RF, LR, DT, SVM, XGB	Public Health Dataset	84%, Support Vector Machine
Siuly *et al.* [45], 2016	MLP, DT, RF, NB, kNN, L-SVM	IoT based generated Data	92.30%, Random Forest and L-SVM
Kumar *et al.* [46], 2021	kNN, RF, SVM	Heart disease dataset, SA	95%, Random Forest
Ammarah *et al.* [47],2021	DT, RF, MLP, kNN, SVM, DL methods, Cross-validation	Data gathered from the hospital	92.9%, Multi-Layer Perceptron

To fill in the gap of accuracy with a standalone classifier, Riyaz et al. [48] have thrown light on the requirement of various ensemble classifiers for the prognosis of heart disease. Edward et al. [49] have used an ensemble classifier by combining a Decision tree, Random Forest, and Extreme gradient boost. They used a majority voting technique in their classifier to achieve a 99.32% better result than the baseline. Different Machine Learning (ML) and deep learning (DL) methods to foresee the risk of CHD using 14 risk factors studied in [50]. Wang et al. [51] have conceptualized a 2-layer stack using combinations of ML classifiers to evaluate the CHD risk. Yekkala et al. [52] have compared the performance of various ensemble classifiers and particle swarm optimization-based methods in CHD prediction. Pattanayak et al. [53] have emphasized the importance of hyperparameter tuning in their research on ensemble ML models and have benchmarked their model efficiency using various standard parameters.

Based on the discussions mentioned above, we learned that researchers have been trying to put forward a number of novel methods to predict the risk associated with the human cardiovascular system using ML classifiers. While individual classifiers are performing well in most of the use cases discussed, there is an even higher upside for employing an ensemble of ML classifiers with logical rules. Another significant aspect of this research is handling the huge amount of data because of which feature engineering plays a pivotal role in the experiment in order to boost the model performance.

III. METHODOLOGY

In our proposed research, we focus on designing a two-step ensemble classifier to forecast the risk of CHD with conventional Machine Learning classifiers as the foundation blocks. Another essential aspect of our model design is the selection of the optimal feature list for the best predictive performance. A high-level architectural flow of our proposed methodology is shown in Figure 1 below.

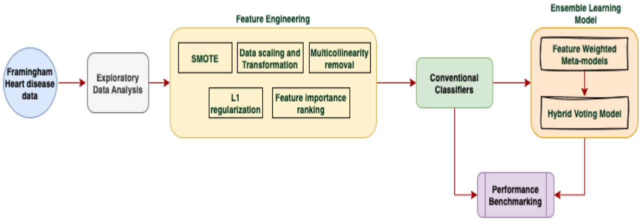

Framingham Heart disease data
Exploratory Data Analysis
Feature Engineering
SMOTE
Data scaling and Transformation
Multicollinearity removal
L1 regularization
Feature importance ranking
Conventional Classifiers
Ensemble Learning Model
Feature Weighted Meta-models
Hybrid Voting Model
Performance Benchmarking

FIGURE 1. Schematic diagram of the proposed methodology.

A. DESCRIPTION OF DATASET

We utilized a dataset gathered from a survey conducted in Framingham, Massachusetts, to examine cardiovascular health. The dataset included 15 clinical and demographic variables for 4,238 participants. However, a small subset of individuals had incomplete information, necessitating the use of imputation to fill in missing values with appropriate data. Table 2 provides a summary of the features that were taken into account for our investigation.

TABLE 2. Description of features.

Feature #	Feature name	Description
Feature 1	male	Gender of the subject; binary data field
Feature 2	age	Age of the subject at the time of data collection
Feature 3	education	Education level of the subject
Feature 4	currentSmoker	If the subject was a smoker at the time of data collection
Feature 5	cigsPerDay	No. of cigarettes consumed per day
Feature 6	BPMeds	If the subject is taking medicines for BP
Feature 7	prevalentStroke	If the subject has a history of stroke
Feature 8	prevalentHyp	If the subject has a history of hypertension
Feature 9	diabetes	If the subject has a history of diabetes
Feature 10	totChol	The Total cholesterol level of the subject
Feature 11	sysBP	The Systolic BP level of the subject
Feature 12	diaBP	The diastolic BP level of the subject
Feature 13	BMI	Body-mass-index of the subject
Feature 14	heartRate	The heart rate of the subject measured randomly
Feature 15	glucose	The Glucose level of the subject measured randomly
Target	TenYearCHD	Risk of developing CHD in 10 years' time frame

The population group covered in our dataset had 85% of subjects who did not have the risk of CHD. The charts in Figure 2 display the distribution of the 15% population who have the risk of developing CHD in 10 years with the risk factors considered to analyze the impact of features on the risk of CHD for the population exposed to the risk.

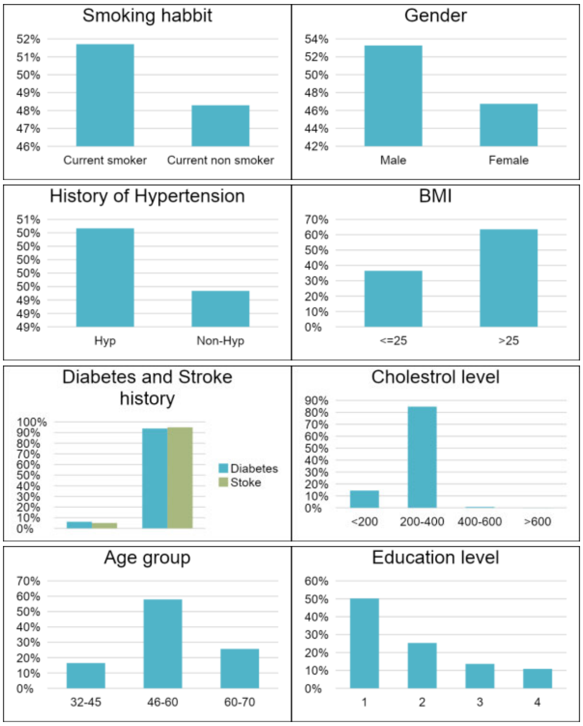


FIGURE 2. Analysis of CHD risk population w.r.t attributes.

We draw the following inference from doing exploratory data analysis on our feature set.

- The male population group is at slightly higher risk as compared to the females
- People with a smoking habit, a history of hypertension, and BMI above the recommended level (19-25), and soaring cholesterol levels have a higher risk of developing CHD
- CHD risk goes down with education level
- CHD risk increases with an increase in age beyond 45
- The target does not have a clear correlation with any one feature

In our data, it is identified that 85% of the population does not have CHD risk as compared to 15% who have the risk of CHD. This skewness towards the absence of CHD risk introduced class imbalance which usually hinders conventional Machine Learning models from performing optimally. Therefore, the Synthetic Minority Oversampling Technique (SMOTE) is used as a data transformation method to mitigate the imbalance. The new input data had 7,188 total observations and an even split between the presence and absence of CHD risk.

B. FEATURE OPTIMIZATION

Feature optimization in Machine Learning refers to the process of choosing the most relevant features from a dataset to use as inputs for a model. This can be done through a variety of methods, such as feature extraction followed by feature selection. The goal is to upgrade the performance of the model by using only the most relevant features. This can also help to reduce overfitting, as well as enhance the computational efficiency of the model.

For the aim of our analysis, the following feature optimization methods are implemented as described in the following subsections.

1) MULTICOLLINEARITY REMOVAL

Multicollinearity is a phenomenon that occurs if the independent variables in a regression model are significantly correlated with each other. Amongst our feature set, strong Pearson Correlation Coefficient (PCC) values are observed between a few features as shown in Table 3.

TABLE 3. **Highly correlated features.**

Features	PCC
sysBP, diaBP	0.78
currentSmoker, cigsPerDay	0.77
prevalentHyp, sysBP	0.70
prevalentHyp, diaBP	0.62
diabetes, glucose	0.61

Based on the input dataset and a chosen statistical level of significance at 95%, all the above PCC values are found to be statistically significant with p-values less than 0.05. The Variance Inflation Factor (VIF) score is used to further assess the aforementioned features since it gauges to what extent the variance of the calculated regression coefficient is inflated by the existence of additional correlated independent variables. Mathematically,

$$VIF_i(f) = 1/(1 - R_i^2) \tag{1}$$

where, $R^2 = \frac{SSR}{SST}$ defined as the coefficient of determination for the regression of feature f on all other independent features denoted by i; SSR: - Sum Square Regression and SST: - Sum Square Total of SSE (Sum Square Error) and SSR

For our dataset, we calculated stepwise VIF scores for each combination of features showing multicollinearity as shown in Table 4.

TABLE 4. **Highly correlated features.**

Step 1		**Step 2**	
Feature	**VIF Score**	**Feature**	**VIF Score**
prevalentHyp	1.795	**prevalentHyp**	1.638
sysBP	**103.766**	diaBP	1.638
diaBP	99.364		
Step 3		**Step 4**	
Feature	**VIF Score**	**Feature**	**VIF Score**
currentSmoker	3.824	**diabetes**	1.11
cigsPerDay	3.824	glucose	1.11

In each phase, the bolded features with the highest VIF score are the ones that are taken out of the dataset. We found that the features have equal VIF scores in subsequent steps, therefore any one of them is eliminated to ensure linear independence.

2) L1 REGULARIZED LOGISTIC REGRESSION

L1 regularization, also known as Lasso regularization, is one of the methods in ML to ward off overfitting by adding a penalty term to the cost function. The penalty term is defined as the absolute value of the coefficients of the model multiplied by a constant, called the regularization parameter. This sets some of the coefficients to zero, effectively minimizing the count of features fed into the model. L1 regularization can be useful for feature selection and preventing overfitting, especially when working with high-dimensional data.

For our dataset, we used an L1-regularized Logistic regression. The rationale for using Logistic regression is that it produces True or False labels for the input features based on their significance of contribution. Formulating the cost function mathematically,

$$W = -(1/N) * \sum_{i=1}^{N} \left[y_i * log(y_i) + (1 - \widehat{y_i}) * \alpha \right] \tag{2}$$

$$\alpha = log(1 - y_i) \tag{3}$$

$$Wnew = W + \lambda * ||W|| \tag{4}$$

where N:- No. of observations, y:- predicted probability, $\hat{y}$:- actual value W:- vector of coefficients; λ:- Regularization parameter; $||W||$:- the L1 norm for W, i.e., the sum of the absolute values of the elements in W.

We observed that all the coefficients of all the 11 features obtained after multicollinearity removal could not be shrunk. This further validated the linear independence of our features.

3) RANDOM FOREST FEATURE IMPORTANCE

Random Forest models have inbuilt feature importance measures that can be used to evaluate the comparative importance of each feature in the dataset for the prediction task. This is achieved by measuring the decrease in impurity (i.e., Gini index) due to splits on a particular feature. It is calculated by averaging the decrease in impurity over all trees in the model forest and is normalized by the total number of samples. The Gini importance of our considered features is given in Figure 3 in descending ranked order.

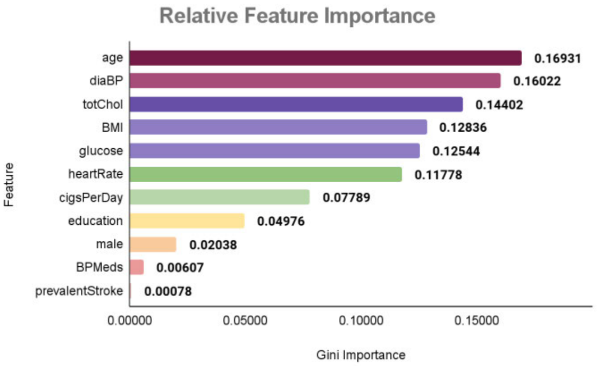


FIGURE 3. **The relative importance of features.**

C. ENSEMBLE LEARNING MODEL

We devised a 2-step ensemble learning model on our features to predict the risk of CHD. A high-level architecture diagram of the proposed model is demonstrated in Figure 4.

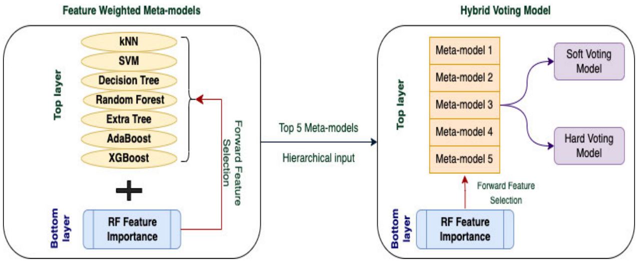


FIGURE 4. **Constituents of the ensemble learning model.**

1) STEP 1 - FEATURE WEIGHTED META-MODELS

The goal of this activity is to construct a two-layer Meta-model that takes the outputs of the lower layer as an input for the layer stacked above it. In our Meta-model, RF Feature Importance forms the lower layer where we have obtained the relative importance of 11 features considered. The Machine Learning classifiers are being used to form the top layer of the model. We are using kNN, SVM, DT, RF, ET, AdB, and XGB classifiers to determine CHD risk. Features from the bottom layer are fed into each classifier in the top layer following an iterative manner with Forward Feature Selection algorithms based on their relative importance, as shown in Figure 3. Since we have removed multicollinearity from our feature set, Forward Feature Selection serves as the most appropriate feature selection method to reduce the computational expense significantly. Hyperparameters of the respective classifiers are optimized with GridSearchCV for the most accurate outcome of CHD risk prediction.

2) STEP 2 - HYBRID VOTING MODEL

With the goal of enhancing the accuracy of our proposed CHD prediction methodology, we are developing the Hybrid Voting ensemble model. The Hybrid voting classifier is a way to enhance the accuracy of the model by combining the predictions of multiple Feature Weighted Meta-Models. The concept behind a voting ensemble classifier is that different models can make different mistakes. However, by combining the predictions of multiple models, the errors can be reduced, leading to improved performance. The voting approach minimizes the errors of individual classifiers and maximizes the prediction space of the model, thus covering the blind spots of each classifier. This helps in decreasing misclassification in predictions, which is particularly critical in CHD risk prediction, where both False Positives and False Negatives have severe consequences. Our hybrid voting ensemble models are less prone to overfitting, as they are less influenced by the idiosyncrasies of a single training set, which makes them well-suited to complex problems with a lot of noise or variance in the data.

Our proposed Hybrid Voting Model takes the 5 best performing Feature Weighted Meta-Models as the constituents to predict the risk of CHD. This Hybrid Voting Model is formed as a hierarchical top layer above the Meta-models constructed in the earlier step. The same Forward Feature Selection algorithm used in the previous step is applied to feed features into the model based on their relative importance. We are considering both Hard and Soft voting models for our evaluation. In hard voting, the final prediction is based on a simple majority vote of the individual model predictions. In soft voting, the final prediction is based on a weighted average of the individual model predictions, where the weights are based on the models' predicted probabilities. Mathematically,

$$Y_{hard} = mode\left[y_i(x)\right], i \in [1, k] \tag{5}$$

$$Y_{soft} = argmax\left\{\sum\nolimits_{i=1}^{k} \left[w_i * y_i(x)\right]\right\} \tag{6}$$

where, Y_{hard} and Y_{soft} are resultant outputs of Hard and Soft Hybrid voting models respectively, y:- prediction of individ-

Algorithm 1 Feature Weighted Meta-Model Using Forward Feature Selection

```
Parameters:
        Current_set = { }
        feature_list: set of all ordered features as per RF Feature
        importance
Initialization:
    Start
    For each Meta-model
      While feature_list is not empty
        Select the next best feature f
        Add the feature f to the current_set
        Remove feature f from feature_list
        Train the Meta-model on the current_set with
        hyperparameter Tuning
        Evaluate the trained hyperparameter tuned model
      End While
    End For
    Sort the Meta-model set by the Performance Parameter and
    consider 5 best performing models
    End
```

Algorithm 2 2-Step Hybrid Ensemble Model

```
Parameters:
        Meta-model = 5
        feature_list: set of all ordered features as per RF Feature
        importance
Initialization:
    Start
    For each Meta-model
      Calculate the Weight of Meta-model
      Append the weights to the Weighted Meta-model
    End For
    While feature_list is not empty
      Select the next best feature f
      Add the feature f to the current_set
      Remove feature f from feature_list
      Calculate hard voting with the Meta-Models
      Calculate soft voting with the Weighted Meta-models
      Hyperparameter Tuning on the Hybrid voting model
      Evaluate the Performance Parameter on both hard voting
      and soft voting
    End While
    Select the feature_list with the best-performing Hybrid voting
    model
    End
```

ual Feature Weighted Meta-Models, w:- weights associated with each Feature Weighted Meta-Model, k:- number of constituent models of the Hybrid voting model

The resultant outcome of the proposed Hybrid Voting Model resonates with the hypothesis provided by the Law of Large Numbers which states that the ensemble model reduces the variance of the predictions compared to a single model, resulting in improved accuracy. The algorithms followed for our proposed model are demonstrated above.

D. MODEL EVALUATION MATRIX

To evaluate the efficiency of our proposed Feature Weighted Meta-Models and Hybrid Voting model, we use the following metrics

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

where TP:- True Positive, TN:- True Negative, FP:- False Positive, and FN:- False Negative

$$\textbf{F1 score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{8}$$

$$\text{where, } Precision = \frac{TP}{TP + FP} \tag{8.a}$$

$$Recall = \frac{TP}{TP + FN} \tag{8.b}$$

The F1 score is calculated as the Harmonic mean of Precision and Recall. F1 Score is a useful metric for our analysis of healthcare data since in the medical domain both False Positive (FP) and False Negative (FN) misclassification instances can be immensely fatal.

$$\textbf{Matthew's Correlation Coefficient (MCC)} = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{9}$$

MCC is an efficiency measure for binary classification and varies in the range of $[-1, 1]$. MCC value 1 indicates perfect alignment, and -1 means total misalignment with the classification. 0 MCC signifies that the classification does not have any logical ground. The level of statistical significance is chosen at 95%.

AUC-ROC: The area captured by the Receiver Operating Characteristic (ROC) curve or AUC-ROC is calculated by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The AUC-ROC is between 0 and 1, where 1 represents a perfect classifier.

$$\textbf{Misclassification Ratio (MCR)} = \frac{FP + FN}{TP + TN + FP + FN} \tag{10}$$

Misclassification Ratio (MCR) can be called the blind spot of the classifier. MCR can also be defined as $[1 - Accuracy]$, therefore describing the error in the classification process. The value of MCR varies between 0 to 1; a smaller MCR indicates better classifier performance.

IV. RESULTS AND ANALYSIS

We conducted two experiments to study the effectiveness of the proposed model. The initial experiment stacked feature optimization on conventional Machine Learning classifiers to build our proposed Feature Weighted Meta-Models and iteratively determined the optimal feature and hyperparameters combination for the best performance. The second experiment selected the top 5 best performing Feature Weighted Meta-Models and designed a Hybrid Voting ensemble classifier to improve CHD risk prediction. All experiments are run on an Intel-Core i5 11th gen. processor @3.7GHz having 8GB of RAM, and a 64-bit Windows 11 operating system using the Python programming language on the Google Colab platform.

A. EXPERIMENT 1: FEATURE WEIGHTED META-MODELS WITH FORWARD FEATURE SELECTION

In this part of the experiment, results from the feature selection method are stacked on conventional classifiers to build our proposed Feature Weighted Meta-Models. The Random Forest Feature Importance used here chalked out features based on their relative importance and ranked them. The Feature Weighted Meta-Models take them as input, following the Forward Feature Selection Algorithm iteratively. These models are trained with manual hyperparameter optimization, and the best results are obtained for 5-fold cross-validation.

Table 5 demonstrates the efficiency of considered stacked classifiers based on Accuracy, F1 Score, MCC (statistically significant, p-value $<$ 0.05), and MCR parameters along with the number of features they considered for best performance.

TABLE 5. Performance of feature weighted meta-models.

Model	No. of Features (#F)	Accuracy (%)	F1	MCC	MCR
kNN	7	84.71	0.87	0.71	0.15
SVM	9	77.31	0.79	0.55	0.23
Decision Tree	6	79.86	0.76	0.51	0.20
Random Forest	8	86.64	0.87	0.72	0.13
Extra Tree	**7**	**89.14**	**0.88**	**0.74**	**0.11**
AdaBoost	7	72.26	0.73	0.45	0.28
XGBoost	8	82.53	0.81	0.61	0.17

As observed in the above table, the Random Forest and Extra Tree models had the highest accuracy at 86.64% and 89.14%, respectively, indicating that they are the best-performing models for correctly classifying instances. These models also had the highest F1 scores, 0.87 and 0.88, respectively, as well indicating that they have a good balance between precision and recall. The Extra Tree model had the highest MCC of 0.74, indicating that it has the best overall performance in terms of true and false classification rates. The Extra Tree, and Random Forest models produced the lowest values of MCR as well, which demonstrates high efficiency in the assignment of correct class labels. Here, we observed that all the Feature Weighted Meta-Models produced the best accuracy at a smaller number of attributes than the original dataset, making the time complexity of our analysis significantly better.

Training and Testing accuracy of various Feature Weighted Meta-Models are shown in Figure 5 which establishes the absence of both significant overfitting and underfitting.

We gauged the performance of our proposed models with AUC-ROC as shown in Figure 6 to measure the capability of a model to differentiate between positive and negative classes. It was observed that all the proposed Feature Weighted Meta-Models performed decently well. However, Random Forest and Extra Tree models had the best AUC-ROC value of 0.94.

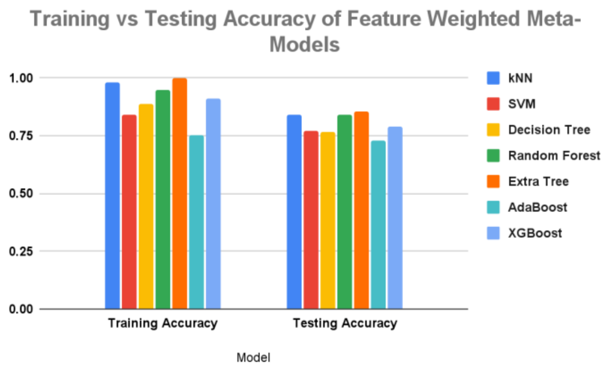


FIGURE 5. Learning accuracy of feature weighted meta-models.

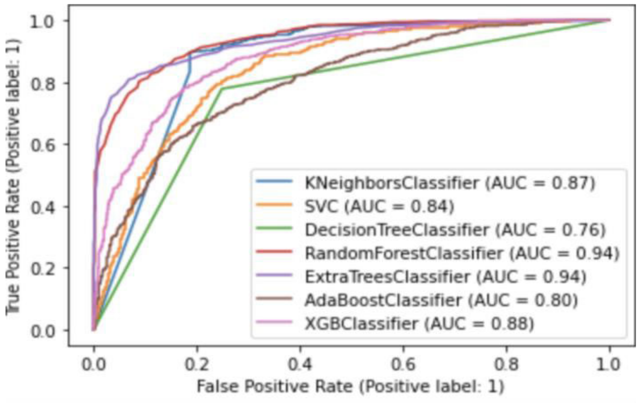


FIGURE 6. ROC curves for the feature weighted meta-models.

The accuracy of the Feature Weighted Meta-Models was benchmarked with that of the conventional classifiers and shown in Figure 7. The conventional classifier is considered without any hyperparameter tuning with all 11 features obtained after eliminating multicollinearity.

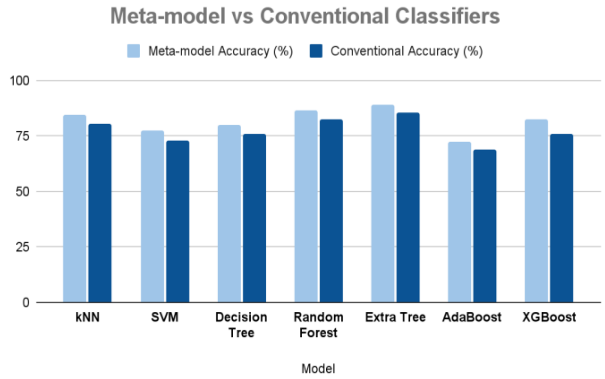


FIGURE 7. Accuracy comparison of feature weighted meta-models and conventional classifiers.

Our proposed Feature Weighted Meta-Models are observed to outperform their respective conventional classifiers.

B. EXPERIMENT 2: HYBRID VOTING MODEL

In the final experiment, we selected the 5 best-performing Feature Weighted Meta-Models from Experiment 1 to design the proposed Hybrid Voting Model. Here, we considered Extra Tree, Random Forest, XGBoost, kNN, and Decision Tree. Forward Feature Selection Algorithm is followed to provide input features to the Hybrid Voting models iteratively. The performances of both Hard and Soft voting-based Hybrid models are shown in Table 6. Both the novel Hybrid Voting models are benchmarked based on Accuracy, F1 Score, MCC, and MCR parameters.

TABLE 6. Hybrid voting model performance.

No. of Features (#F)	Accuracy		F1 Score		MCC		MCR	
	Soft	Hard	Soft	Hard	Soft	Hard	Soft	Hard
1	63.84	64.39	0.66	0.67	0.28	0.29	0.36	0.36
2	71.77	70.86	0.70	0.68	0.44	0.43	0.28	0.29
3	74.97	74.41	0.76	0.75	0.50	0.49	0.25	0.26
4	76.91	77.61	0.78	0.78	0.54	0.55	0.23	0.22
5	81.29	82.41	0.82	0.83	0.63	0.65	0.19	0.18
6	87.17	84.91	0.84	0.85	0.66	0.70	0.13	0.15
7	**95.87**	**87.00**	**0.91**	**0.87**	**0.83**	**0.74**	**0.04**	**0.13**
8	88.19	85.88	0.86	0.86	0.71	0.72	0.12	0.14
9	90.98	85.33	0.86	0.86	0.70	0.71	0.09	0.15
10	87.70	85.54	0.86	0.86	0.70	0.71	0.12	0.14
11	84.01	84.98	0.85	0.85	0.68	0.70	0.16	0.15

The Hard and Soft Hybrid Voting models showed AUC-ROC values of 0.86 and 0.94 respectively as shown in Figure 8.

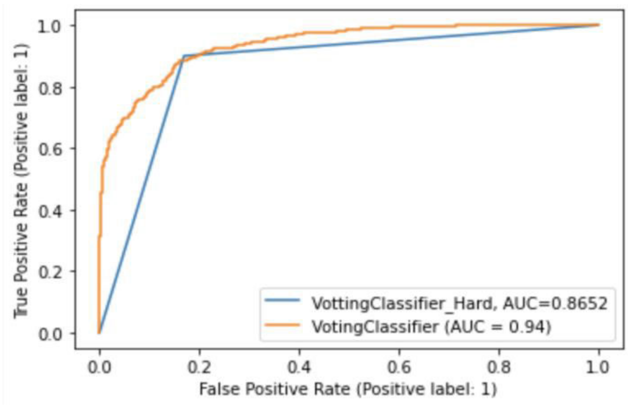


FIGURE 8. ROC curves for the hybrid voting models.

We are using Analysis of Variance (ANOVA) to establish the statistical significance of the intra and inter-group variations among the output of Meta-models as well as the Hybrid voting models. The Results of a One-way ANOVA performed with the observation matrix constructed with the proposed models' predicted target value are shown in Table 7.

TABLE 7. ANOVA table of model performance.

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	55.338	8	6.917	52.818	7.48E-85	1.939
Within Groups	1693.761	12933	0.131			
Total	1749.099	12941				

SS: Sum of Square, df: Degree of Freedom, MS: Mean Square, F: F-ratio, F crit: Critical value of F-ratio.

As observed, the calculated F-statistic is noticeably higher than the critical value and, the obtained p-value is significantly smaller than the chosen significance level (0.05). Hence, we reject the null hypothesis. This proves that the mean accuracies obtained from the various models are not the same. Since the accuracy obtained from the hybrid voting models (refer to Table 6) is more than that of individual Meta-models (refer to Table 5), we infer that the hybrid models' performance is far superior to that of the individual models.

Figure 9 and Figure 10 illustrate a comparison of Accuracy and MCR for both the Hybrid Voting classifiers with Hard and Soft voting techniques, further substantiating that soft voting performs better than the Hard voting model as the Accuracy improves. We also observed that MCR for the Soft voting model is significantly less than the Hard voting model, indicating better blind spot reduction for an effective prediction improvement.

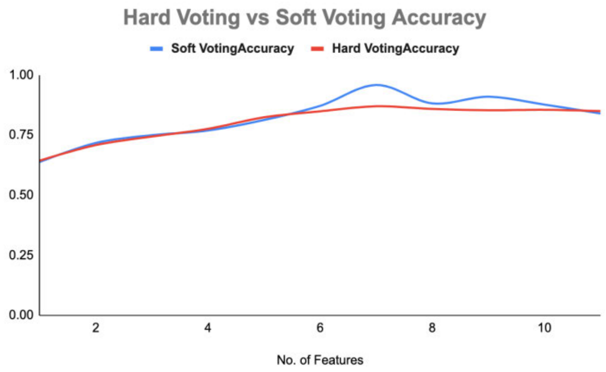


FIGURE 9. Accuracy comparison of hybrid voting.

As observed in our experiments, the proposed Hybrid Soft Voting Model is able to predict the risk failure with 95.87% accuracy by considering only 7 features: age, diaBP, totChol, BMI, glucose, heartRate, and cigsPerDay. This shows a

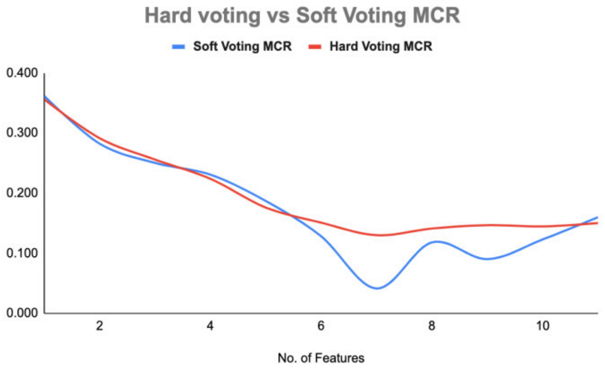


FIGURE 10. MCR comparison of hybrid voting.

23.94% improvement over the conventional classifier baseline and a 17.23% improvement over the Feature Weighted Meta-Models baseline accuracy parameters. The features considered for this prediction result are age, diaBP, totChol, BMI, glucose, heartRate, and cigsPerDay. This model also demonstrated a 0.91 F1 score indicating effective mitigation of risk due to False positive and False negative misclassification. The F1 score improved by 12% over the conventional classifier baseline F1. The MCR of the model is curtailed by 62.25% with a soft voting approach as compared to the best-performing Feature Weighted Meta-model with an Extra Tree classifier. The newly designed Hybrid Hard Voting Model produced a prediction accuracy of 87.00% and an F1 score of 0.87 for the same 7 features. Figure 11 shows a significant decline in MCR for the Hybrid Soft Voting Model as compared to the constituent Meta-models, which indicates that the Hybrid Voting Model was able to minimize the blind spots of individual Meta-models effectively to produce a better prediction for CHD risk. The plot in Figure 12 illustrates that the training and testing accuracy of both hard and soft Hybrid Voting Models are in order and it indicates an absence of the risk of overfitting or underfitting.

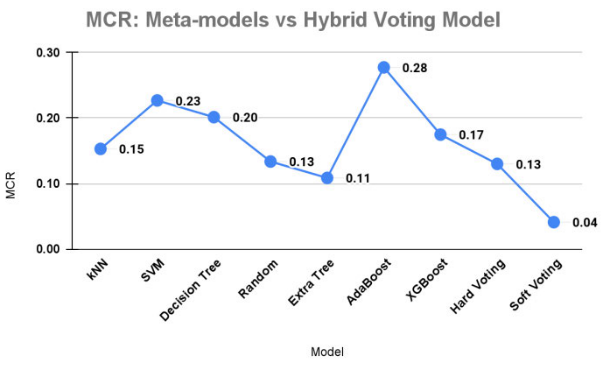


FIGURE 11. Comparison of MCR for all models.

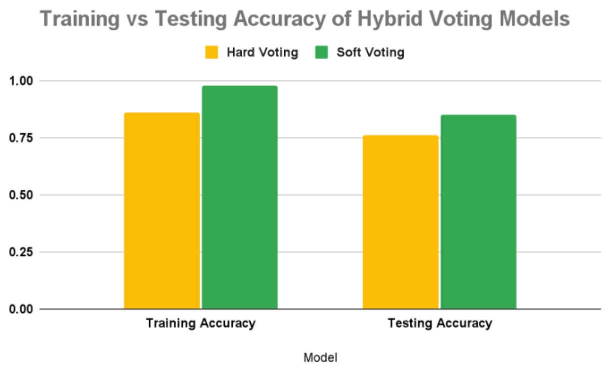


FIGURE 12. Learning accuracy of hybrid voting models.

Therefore, it is evident that our proposed Hybrid Soft Voting model outperforms all other conventional models, Feature Weighted Meta-Models, and Hybrid Hard Voting Model in predicting the risk of CHD.

V. CONCLUSION AND FUTURE SCOPE

In this study, we developed a robust Hybrid Voting Ensemble learning model for predicting the risk of CHD more effectively than conventional classifiers. Our approach included optimizing the features in our Framingham Heart disease dataset and determining their importance towards the outcome in the first stage. Next, we integrated the ranked feature model with conventional classifiers to construct Feature Weighted Meta Models using the Forward Feature Selection algorithm. Finally, we selected the top 5 performing Feature Weighted Meta-Models to form the proposed Hybrid Voting Model. This model achieved an accuracy of 95.87% in predicting the risk of CHD over a ten-year period, as observed in the source dataset. Additionally, we emphasized the importance of the F1 score as a critical metric, as misclassification can have significant consequences in the healthcare domain. Our proposed model demonstrated a high F1 score of 0.91, significantly reducing the risk of misclassification. We also established that our proposed model is able to produce superior predictive performance with 7 features, which are age, diaBP, totChol, BMI, glucose, heartRate, and cigsPerDay. Furthermore, the newly developed models performed better with fewer features, indicating optimized time complexity. Therefore, our proposed Hybrid Voting Model can aid healthcare practitioners in making more accurate predictions about the risk of CHD over a longer timeframe.

Going forward, we plan to incorporate additional datasets to enhance the reliability of our conclusions. We shall also employ metaheuristic techniques and nature-inspired algorithms to ameliorate the parameters of ML classifiers and DL methods for a more efficient evaluation of heart disease across various heart disease-related datasets. Our goal is to improve the accuracy of current algorithms and identify new insights.

ACKNOWLEDGMENT

The authors would like to thank the Department of Science and Technology (DST), Government of India, for their fellowship support during the doctoral program of Sushree Chinmayee Patra and also would like to thank Amrita Vishwa Vidyapeetham for facilitating the research.

CONFLICT OF INTEREST

The authors hereby declare the absence of any competing interest.

REFERENCES

[1] C. Fryar, T.-C. Chen, and X. Li, "Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999–2010," *NCHS Data Brief*, vol. 103, pp. 1–8, Aug. 2012.

[2] WHO. *The Top 10 Causes of Death*. Accessed: Dec. 30, 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[3] L. A. Allen, L. W. Stevenson, K. L. Grady, N. E. Goldstein, D. D. Matlock, R. M. Arnold, N. R. Cook, G. M. Felker, G. S. Francis, P. J. Hauptman, and E. P. Havranek, "Decision making in advanced heart failure: A scientific statement from the American Heart Association," *Circulation*, vol. 125, pp. 1928–1952, Apr. 2012.

[4] R. K. Kodali, G. Swamy, and B. Lakshmi, "An implementation of IoT for healthcare," in *Proc. IEEE Recent Adv. Intell. Comput. Syst. (RAICS)*, Dec. 2015, pp. 411–416.

[5] K. C. Kavitha and R. Perumalraja, "Smart wireless healthcare monitoring for drivers community," in *Proc. Int. Conf. Commun. Signal Process.*, Peru, Apr. 2014, pp. 1105–1108.

[6] D. Goyal, J. Bhaskar, and P. Singh, "Designing the low cost patient monitoring device (LCPMD) & ubiquitous based remote health monitoring and health management system using tablet PC," in *Proc. 2nd IEEE Int. Conf. Parallel, Distrib. Grid Comput.*, Dec. 2012, pp. 7–11.

[7] S. C. Patra and M. R. Kabat, "Placement of relay nodes in WBAN improved by Free Search Krill Herd optimization," in *Proc. 6th Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, Mar. 2019, pp. 654–659.

[8] T. K. Samal, S. C. Patra, and M. R. Kabat, "An adaptive cuckoo search based algorithm for placement of relay nodes in wireless body area networks," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 5, pp. 1845–1856, May 2022.

[9] H. Bhatia, S. N. Panda, and D. Nagpal, "Internet of Things and its applications in healthcare—A survey," in *Proc. 8th Int. Conf. Rel., INFOCOM Technol. Optim. Trends Future Directions (ICRITO)*, Jun. 2020, pp. 305–310.

[10] D. Sahu, B. Pradhan, A. Khasnobish, S. Verma, D. Kim, and K. Pal, "The Internet of Things in geriatric healthcare," *J. Healthcare Eng.*, vol. 2021, Jul. 2021, Art. no. 6611366.

[11] P. K. Fahad and M. S. Pallavi, "Prediction of human health using machine learning and big data," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Chennai, India, Apr. 2018, pp. 0195–0199, doi: 10.1109/ICCSP.2018.8524352.

[12] N. J. Ravindran and P. Gopalakrishnan, "Predictive analysis for healthcare sector using big data technology," in *Proc. 2nd Int. Conf. Green Comput. Internet Things (ICGCIoT)*, Bangalore, India, Aug. 2018, pp. 326–331, doi: 10.1109/ICGCIoT.2018.8753090.

[13] T. K. Ramesh and A. Shashikanth, "A machine learning based ensemble approach for predictive analysis of healthcare data," in *Proc. 2nd PhD Colloq. Ethically Driven Innov. Technol. Soc. (PhD EDITS)*, Bangalore, India, Nov. 2020, pp. 1–2, doi: 10.1109/PhDEDITS51180.2020.9315300.

[14] S. Kusuma and J. D. Udayan, "Analysis on deep learning methods for ECG based cardiovascular disease prediction," *Scalable Comput., Pract. Exper.*, vol. 21, no. 1, pp. 127–136, Mar. 2020.

[15] V. Viswanathan, A. D. Jamthikar, D. Gupta, A. Puvvula, N. N. Khanna, L. Saba, K. Viskovic, S. Mavrogeni, J. R. Laird, G. Pareek, M. Miner, P. P. Sfikakis, A. Protogerou, A. Sharma, P. Kancharana, D. P. Misra, V. Agarwal, G. D. Kitas, A. Nicolaides, and J. S. Suri, "Does the carotid bulb offer a better 10-year CVD/stroke risk assessment compared to the common carotid artery? A 1516 ultrasound scan study," *Angiology*, vol. 71, no. 10, pp. 920–933, Nov. 2020, doi: 10.1177/0003319720941730.

[16] N. N. Khanna, A. D. Jamthikar, D. Gupta, T. Araki, M. Piga, L. Saba, C. Carcassi, A. Nicolaides, J. R. Laird, H. S. Suri, A. Gupta, S. Mavrogeni, A. Protogerou, P. Sfikakis, G. D. Kitas, and J. S. Suri, "Effect of carotid image-based phenotypes on cardiovascular risk calculator: AECRS1.0," *Med. Biol. Eng. Comput.*, vol. 57, no. 7, pp. 1553–1566, Jul. 2019, doi: 10.1007/s11517-019-01975-2.

[17] F. Ali, S. El-Sappagh, S. M. R. Islam, D. Kwak, A. Ali, M. Imran, and K.-S. Kwak, "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion," *Inf. Fusion*, vol. 63, pp. 208–222, Nov. 2020.

[18] N. N. Khanna, A. D. Jamthikar, T. Araki, D. Gupta, M. Piga, L. Saba, C. Carcassi, A. Nicolaides, J. R. Laird, H. S. Suri, A. Gupta, S. Mavrogeni, G. D. Kitas, and J. S. Suri, "Nonlinear model for the carotid artery disease 10-year risk prediction by fusing conventional cardiovascular factors to carotid ultrasound image phenotypes: A Japanese diabetes cohort study," *Echocardiography*, vol. 36, no. 2, pp. 345–361, Feb. 2019.

[19] P. K. Jain, K. V. Tadepalli, S. Roy, and N. Sharma, "Exploring deep learning for carotid artery plaque segmentation: Atherosclerosis to cardiovascular risk biomarkers," *Multimedia Tools Appl.*, pp. 1–33, Oct. 2023, doi: 10.1007/s11042-023-17243-3.

[20] Y. Muhammad, M. Tahir, M. Hayat, and K. T. Chong, "Early and accurate detection and diagnosis of heart disease using intelligent computational model," *Sci. Rep.*, vol. 10, no. 1, pp. 1–17, Nov. 2020.

[21] A. Junejo, Y. Shen, A. A. Laghari, X. Zhang, and H. Luo, "Notice of retraction: Molecular diagnostic and using deep learning techniques for predict functional recovery of patients treated of cardiovascular disease," *IEEE Access*, vol. 7, pp. 120315–120325, 2019.

[22] F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein, "Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features," in *Proc. Int. Workshop Stat. Atlases Comput. Models Heart*. Cham, Switzerland: Springer, 2017, pp. 120–129.

[23] S. Wongvibulsin, S. S. Martin, S. R. Steinhubl, and E. D. Muse, "Connected health technology for cardiovascular disease prevention and management," *Current Treat. Options Cardiovascular Med.*, vol. 21, no. 6, p. 29, Jun. 2019, doi: 10.1007/s11936-019-0729-0.

[24] N. Alshurafa, C. Sideris, M. Pourhomayoun, H. Kalantarian, M. Sarrafzadeh, and J.-A. Eastwood, "Remote health monitoring outcome success prediction using baseline and first month intervention data," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 2, pp. 507–514, Mar. 2017, doi: 10.1109/JBHI.2016.2518673.

[25] M. A. Karaolis, J. A. Moutiris, D. Hadjipanayi, and C. S. Pattichis, "Assessment of the risk factors of coronary heart events based on data mining with decision trees," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 3, pp. 559–566, May 2010, doi: 10.1109/TITB.2009.2038906.

[26] V. S. H. Rao and M. N. Kumar, "Novel approaches for predicting risk factors of atherosclerosis," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 1, pp. 183–189, Jan. 2013, doi: 10.1109/TITB.2012.2227271.

[27] M. Giardina, F. Azuaje, P. McCullagh, and R. Harper, "A supervised learning approach to predicting coronary heart disease complications in type 2 diabetes mellitus patients," in *Proc. 6th IEEE Symp. Bioinf. Bioeng. (BIBE)*, Oct. 2006, pp. 325–331, doi: 10.1109/bibe.2006.253297.

[28] R. Chen, Y. Yang, F. Miao, Y. Cai, D. Lin, J. Zheng, and Y. Li, "3-year risk prediction of coronary heart disease in hypertension patients: A preliminary study," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2017, pp. 1182–1185, doi: 10.1109/EMBC.2017.8037041.

[29] D. Krishnani, A. Kumari, A. Dewangan, A. Singh, and N. S. Naik, "Prediction of coronary heart disease using supervised machine learning algorithms," in *Proc. IEEE Region 10 Conf. (TENCON)*, Oct. 2019, pp. 367–372, doi: 10.1109/TENCON.2019.8929434.

[30] T. A. Assegie, P. K. Rangarajan, N. K. Kumar, and D. Vigneswari, "An empirical study on machine learning algorithms for heart disease prediction," *IAES Int. J. Artif. Intell.*, vol. 11, no. 3, p. 1066, Sep. 2022.

[31] S. Subbulakshmi and K. V. Adarsh, "Systematic cardiovascular disorder identification using machine learning algorithms," *Specialusis Ugdymas*, vol. 1, no. 43, pp. 8615–8627, 2022.

[32] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker, "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020, doi: 10.1109/access.2020.2980942.

[33] B. Nithya and V. Ilango, "Predictive analytics in health care using machine learning tools and techniques," in *Proc. Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, Jun. 2017, pp. 492–499, doi: 10.1109/ICCONS.2017.8250771.

[34] C. A. U. Hassan, J. Iqbal, R. Irfan, S. Hussain, A. D. Algarni, S. S. H. Bukhari, N. Alturki, and S. S. Ullah, "Effectively predicting the presence of coronary heart disease using machine learning classifiers," *Sensors*, vol. 22, no. 19, p. 7227, Sep. 2022, doi: 10.3390/s22197227.

[35] C. Bemando, E. Miranda, and M. Aryuni, "Machine-learning-based prediction models of coronary heart disease using Naïve Bayes and random forest algorithms," in *Proc. Int. Conf. Softw. Eng. Comput. Syst., 4th Int. Conf. Comput. Sci. Inf. Manage. (ICSECS-ICOCSIM)*, Aug. 2021, pp. 232–237, doi: 10.1109/ICSECS52883.2021.00049.

[36] K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, "Multiple disease prediction using machine learning algorithms," *Mater. Today Proc.*, vol. 80, pp. 3682–3685, Jan. 2021.

[37] 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, Oct. 2021, Art. no. 012013.

[38] V. T. Truong, B. P. Nguyen, T.-H. Nguyen-Vo, W. Mazur, E. S. Chung, C. Palmer, J. T. Tretter, T. Alsaied, V. T. Pham, H. Q. Do, P. T. N. Do, V. N. Pham, B. N. Ha, H. N. Chau, and T. K. Le, "Application of machine learning in screening for congenital heart diseases using fetal echocardiography," *Int. J. Cardiovascular Imag.*, vol. 38, no. 5, pp. 1007–1015, May 2022.

[39] A. S. Abdalrada, J. Abawajy, T. Al-Quraishi, and S. M. S. Islam, "Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: A retrospective cohort study," *J. Diabetes Metabolic Disorders*, vol. 21, no. 1, pp. 251–261, Jan. 2022.

[40] N. Singh and S. Bhatnagar, "Machine learning for prediction of drug targets in microbe associated cardiovascular diseases by incorporating host-pathogen interaction network parameters," *Mol. Informat.*, vol. 41, no. 3, Mar. 2022, Art. no. 2100115.

[41] B. P. Doppala, D. Bhattacharyya, M. Janarthanan, and N. Baik, "A reliable machine intelligence model for accurate identification of cardiovascular diseases using ensemble techniques," *J. Healthcare Eng.*, vol. 2022, Mar. 2022, Art. no. 2585235.

[42] E. F. Gudmundsson, G. Björnsdottir, S. Sigurdsson, K. Andersen, B. Thorsson, T. Aspelund, and V. Gudnason, "Carotid plaque is strongly associated with coronary artery calcium and predicts incident coronary heart disease in a population-based cohort," *Atherosclerosis*, vol. 346, pp. 117–123, Apr. 2022.

[43] A. Kishor and W. Jeberson, "Diagnosis of heart disease using Internet of Things and machine learning algorithms," in *Proc. 2nd Int. Conf. Comput., Commun., Cyber-Secur.*, Ghaziabad, India, Oct. 2020, pp. 691–702.

[44] L. Marco and G. M. Farinella, *Computer Vision for Assistive Healthcare*. Cambridge, MA, USA: Academic Press, 2018.

[45] S. Siuly and Y. Zhang, "Medical big data: Neurological diseases diagnosis through medical data analysis," *Data Sci. Eng.*, vol. 1, no. 2, pp. 54–64, Jun. 2016.

[46] D. Kumar, C. Verma, A. Gupta, M. S. Raboaca, and B. Bakariya, "Detection of cardiac disease and association with family history using machine learning," in *Proc. 10th Int. Conf. Syst. Model. Advancement Res. Trends (SMART)*, Dec. 2021, pp. 670–675, doi: 10.1109/SMART52563.2021.9676234.

[47] Umm-e-Ammarah, F. Bukhari, M. Idrees, and W. Iqbal, "Predictive analysis of congenital heart defects prior to birth," in *Proc. Int. Conf. Robot. Autom. Ind. (ICRAI)*, Oct. 2021, pp. 1–6, doi: 10.1109/ICRAI54018.2021.9651436.

[48] L. Riyaz, M. A. Butt, and M. Zaman, "Ensemble learning for coronary heart disease prediction," in *Proc. 2nd Int. Conf. Intell. Technol. (CONIT)*, Jun. 2022, pp. 1–9, doi: 10.1109/CONIT55038.2022.9848292.

[49] J. Edward, M. M. Rosli, Y.-A. Chua, N. A. M. Kasim, and H. Nawawi, "Classification prediction of familial hypercholesterolemia using ensemble-based classifier with feature selection and rebalancing technique," in *Proc. 13th Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2022, pp. 278–283, doi: 10.1109/ICTC55196.2022.9952820.

[50] 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, Jul. 2021, Art. no. 8387680.

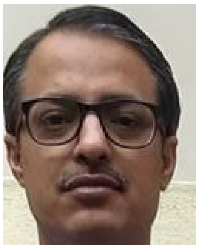
[51] J. Wang, C. Liu, L. Li, W. Li, L. Yao, H. Li, and H. Zhang, "A stacking-based model for non-invasive detection of coronary heart disease," *IEEE Access*, vol. 8, pp. 37124–37133, 2020, doi: 10.1109/ACCESS.2020.2975377.

[52] I. Yekkala, S. Dixit, and M. A. Jabbar, "Prediction of heart disease using ensemble learning and particle swarm optimization," in *Proc. Int. Conf. Smart Technol. Smart Nation (SmartTechCon)*, Aug. 2017, pp. 691–698, doi: 10.1109/SmartTechCon.2017.8358460.

[53] S. Pattanayak and T. Singh, "Cardiovascular disease classification based on machine learning algorithms using GridSearchCV, cross validation and stacked ensemble methods," in *Proc. 6th Int. Conf. Adv. Comput. Data Sci. (ICACDS)*, Kurnool, India. Cham, Switzerland: Springer, Apr. 2022, pp. 219–230.

SUSHREE CHINMAYEE PATRA received the B.Tech. degree in CSE from the Biju Patnaik University of Technology (BPUT), Odisha, India, and the M.Tech. degree in CSE from VSSUT, Burla, Odisha. She is currently pursuing the Ph.D. degree with the Department of CSE, Amrita Vishwa Vidyapeetham, Bengaluru, India. Her research interests include artificial intelligence, machine learning, the IoT, and bioinformatics. She was a recipient of the INSPIRE Fellowship conferred by the Department of Science and Technology (DST), Government of India. She has also received the Governor's Gold Medal for Academic Excellence for securing the First position in the university during the M.Tech. Program.

B. UMA MAHESWARI (Senior Member, IEEE) received the B.E. degree in computer science and engineering from Bharathidasan University, in 1993, the M.E. degree in computer science and engineering from Anna University, Chennai, India, in 2004, and the Ph.D. degree in computer science and engineering from Amrita Vishwa Vidyapeetham, India, in 2020. She is currently an Assistant Professor with the Department of Computer Science and Engineering, Amrita School of Computing, Bengaluru, India. Her current research interests include digital twin, machine/deep learning, the IoT applications in healthcare, agriculture and network security, overlay networks, P2P video streaming, application layer multicasting in wired, and wireless networks.



PEETA BASA PATI (Senior Member, IEEE) received the M.Sc. (Eng.) and Ph.D. degrees from the Indian Institute of Science, Bangalore. He is currently a Professor with the Department of CSE, Amrita Vishwa Vidyapeetham, Bengaluru. His research interests include document digitization and information capture. He has close to 25 years of experience in this field which includes more than 15 years of industrial experience. Prior to joining Amrita Vishwa Vidyapeetham, he was with Cognizant Technology Solutions as a Chief Architect. In this role, he has built and managed IDP systems and implemented and successfully productized IDP systems for multiple business domains and organizations. As part of this, he has experience in dealing with documents that are structured and unstructured, typed written and handwritten, dealing with documents with graphical information content. He is also an Alumnus of NIT Rourkela. He has multiple publications and patents to his credit. He has delivered multiple tutorials in the document processing area and has served in the program and technical committees of multiple national and international conferences, besides being a reviewer of papers for conferences and journals.