International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys.

http://dx.doi.org/10.12785/ijcds/XXXXXX

Cardiovascular disease prediction: Ranking the significant features using Hybrid Accumulated Feature Selection (HAFS) Method

D Yaso Omkari[1], Snehal Shinde[2], Tausif Diwan *[2], Nileshchandra K Pikle[2] and Pradnya Borkar *[3]

[1]*Vellore Institute of Technology, AP Campus, India.*
[2]*Computer Science and Engineering, Indian Institute of Information Technology, Nagpur, India.*
[3]*Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, India.*

Received , Revised , Accepted , Published

Abstract: Heart disease has become a major problem recently, lowering people's standard of living. There is a pressing need to improve prediction models for cardiac data, machine learning has achieved outstanding results in predicting and decision-making. To test the suggested model, this research makes use of the heart disease dataset, which has more than 70,000 records. Body Mass Index, Mean Arterial Pressure, and Pulse Pressure are three additional features that have been enhanced to the dataset in order to enhance the performance. For the most important feature selection, this research suggests the HAFS (Hybrid Accumulated Feature Selection) model. The HAFS design incorporates three statistical methods: Mutual Information (MI), the ANOVA f-test, and the Chi-squared test. The investigation is conducted with the use of various ML and DL classification algorithms, including SVM, NB, LR, XGBoost, LGBoost, AdaBoost, Stochastic gradient descent, and ANN. The experimental findings show that ANN and XGBoost are the best.This work highlights the crucial importance of feature engineering and hyperparameter adjustment in enhancing the accuracy of predictive models.These findings support the ongoing endeavours to create dependable and efficient instruments for the early identification and intervention of cardiac disease.Investigating advanced feature selection techniques and hyperparameter optimization methods can further enhance model performance.

Keywords: Heart Disease, Machine Learning Models, Artificial Neural Networks, Feature Engineering.

1. INTRODUCTION

Many people's habits have changed, for better or worse, as a result of technological advancements in the last several decades, and these shifts have an impact on people's health. People are at risk of heart disease and other health problems due to a decline in physical activity and an increase in their time spent online. Cardiovascular diseases are receiving significant attention [1]. Health records contain valuable information about the behavioural patterns that contribute to these diseases [2], [3]. Before diagnosing heart disease, several tests are typically conducted, including auscultation, blood pressure measurement, cholesterol levels, electrocardiograms, and blood sugar analysis. The results of these tests guide the prescription of appropriate medications [4]. One area where healthcare technology is constantly improving is the detection of cardiovascular problems using Machine Learning (ML) [5], [6]. ML has enormous promise for improving healthcare by deriving novel and substantial insights from the massive volumes of data generated every day by the healthcare industry. The majority of risk prediction algorithms concentrate on a limited number of risk factors. As a result of complex interactions between risk factors, these prediction systems struggle to perform well. Utilizing ML classifiers for data processing can greatly assist in predicting cardiac conditions [7], [8], [9].

Numerous studies have investigated the application of Machine Learning techniques to accurately diagnose cardiac diseases [10], [11], [12]. An effective disease prediction may be hindered by choosing relevant features [13], a limited number of medical datasets, and a lack of in-depth analysis of risk factors. In most cases, these models are trained and evaluated using datasets that are publicly available. Patients' disease status and associated risk factors are included in these datasets. Datasets from Kaggle and the UCI machine learning library were most often employed in the experiments. Coronary Artery Disease (CAD) identification is best accomplished with the Cleveland dataset [14], [15], the UCI heart disease dataset [16], and the Z-Alizadeh Sani dataset [17]. Despite the fact that medical data is generated in massive quantities in the actual world, many studies are utilizing limited datasets. Increasing the dataset typically results in a decline in the ML model's performance. In order to solve this problem, the best ML models must be

E-mail address: sshinde@iiitn.ac.in, tdiwan@iiitn.ac.in, npikle@iiitn.ac.in, pradnyaborkar2@gmail.com

found using large and up-to-date databases. Our data set for this study comes from Kaggle's extensive heart disease dataset, which includes more than 70,000 entries. Therefore, creating a system that improves diagnosis by combining knowledge and experience is the main goal of this article. The main goal of this study is to propose a cardiovascular disease prediction system that utilizes several ML and dDL methods to get highly accurate outcomes. In this study, we utilized a substantial cardiovascular dataset to evaluate various classification techniques, including XGBoost, LGBoost, AdaBoost, Stochastic Gradient Descent (SGD), Logistic Regression (LR), Support Vector Machine (SVM), and Artificial Neural Networks (ANN). Three different ranking approaches, including the ANOVA f-test, chi-squared test, and Mutual Information, were employed to choose features.

The following are the proposed objectives:

1) Use a large dataset with 70,000 examples obtained from Kaggle to examine several ML and DL classification methods.
2) To enhance disease prediction, it is important to investigate the main risk factors for cardiovascular disease. The result is an expansion of the current dataset to include new metrics including BMI, MAP, and PP, or pulse pressure.
3) To determine which features are most crucial by utilizing the ANOVA f-test, chi-squared test, and Mutual Information (MI) feature selection approaches.
4) To determine the best ML and DL models for predicting cardiovascular risk.
5) To incorporate a method for the accurate and efficient prediction of cardiovascular disease.

The paper is organized as follows: Section II provides an overview of related studies on the cardiovascular disease dataset relevant to the problem statement. Section III details the dataset description, feature selection methods, and various ML and DL techniques used for heart disease classification. Section IV discusses the proposed HAFS (Hybrid Accumulated Feature Selection) model's framework and the various metrics for evaluating system performance. Section V presents the preprocessing steps and a comparative analysis of the experimental results obtained in this work. Section VI concludes with the findings derived from the results.

2. Literature Review

The research that are presented in this literature review are those that are in line with our problem description. The several risk variables for the prediction of heart disease were examined by [18] in their 2019 comparison study. The data was normalized using min-max scalar and "k-fold cross-validation" was used in this investigation and trained using RF, KNN, LR, and NB. It has been noted that RF outperforms all algorithms. However, in order to improve the outcomes, this study should be supplemented with several feature selection approaches. Some of the studies focused more on feature selection approaches to get more reliable results. "Embedded feature selection", "filter methods", and "wrapper methods" were the three feature extraction strategies that were compared in [19]. They used RF, SVM, KNN, NB, and XGBoost algorithms to assess feature subset performance after acquiring them using these methods. With the limited set of features, XGBoost performed better than any other algorithm. Base classifiers were compared using ensemble modelling approaches in a model that was built by Shorewala et al. [20]. In this work, the Pearson coefficient was used to examine feature correlations. For the purpose of feature selection, the "Least Absolute Scalp and Selection Operator (LASSO)" was utilized. A thorough examination was conducted using bagging, boosting, and stacking methods. When compared to base models, stacking emerges as the clear winner. Parameter adjustment and other cross-validation techniques can further enhance the model's performance. One more study that compared the effectiveness of base models with ensemble strategies like as stacking, bagging, and boosting demonstrated that ensemble techniques are effective [21]. The LASSO approach is utilized to get the best features. The authors discovered that bagging models improve accuracy by approximately 2

In order to find patterns in massive datasets, many writers use Data Mining, which combines ML, statistics, and database systems. According to Martins et al. (2021) [22], the following five basic classifiers were used: decision tree (DT), optimized decision tree, rule induction, DL, and RF. The optimized decision tree had the best overall performance during the study. Additional research utilizing cross-validation methods is, nevertheless, required.

Implementing cross-validation into the model will yield the best results. A number of research conducted in-depth analyses of ML and DL model performance on both small and large datasets. The authors of [23] utilized two datasets: one from Kaggle's heart disease (70000 records) and another from Cleveland (303 records). The results of several cross-validation methods—including "hold-out, k-fold, repeated random, and stratified k-fold" were examined in this research. The most accurate neural networks were those that used hold-out cross-validation on the Kaggle dataset and the RF networks that used the repeated random approach on the Cleveland dataset. Another study [24] utilized two datasets, one being the UCI arrhythmia dataset (452 records) and the other being the Kaggle heart disease dataset (70000 records). After applying RF, Extra trees, Gradient boosting, and bagging algorithms to the massive dataset, the two techniques with the best accuracy were determined to be RF and Gradient boosting. Parameter optimization is necessary for this investigation, however all ensemble techniques work well on small datasets. A thorough evaluation of the dataset's risk variables is crucial to improve the model's performance with a huge dataset. Several feature selection methods, including ANOVA f-test and f-classify procedures, were utilized in [25]. Using the top3,8, and 12 features, the authors of this study evaluated the efficiency of ML algorithms. Out of all the ML methods, SVM using the top three features and RF using the top eight and twelve features both achieved the best accuracy. The writers of this study solely focused on feature selection.

Selecting the optimal features is crucial for enhancing results. Parameter optimization significantly improves model performance. In the proposed work to achieve better results, new features are generated using feature engineering techniques. With the help of feature engineering and hyperparameter tuning the model performance is enhanced.

3. Materials and Methods

ML models rely heavily on feature selection or extraction for their pattern recognition. In general, the large data decreases prediction accuracy, and also not all features are crucial to detect the label of the data class [26]. This section describes the dataset, feature selection method, and various classification models used for the classification of heart disease.

A. *Materials*

Important information regarding the dataset, including its size, origin, and pertinent aspects, is presented in this subsection.

1) Dataset

The Heart Disease Dataset of 70,000 patient records with 12 attributes, was particularly obtained from Kaggle and used in this research. All of these factors add together to determine how likely someone is to suffer from heart disease. There were three main categories of features found in the dataset:

1) Objective features include patient-related information such as age, height, weight, and gender.
2) Examination features encompass data obtained from medical examination results.
3) Subjective features consist of information of the patient about habits and personal history.

All of the attributes and the types of values for them are described in great depth inTable I.

B. *Feature selection methods:*

1) ANOVA f-test

A statistical technique called "ANalysis Of VAriance" is utilized to assess the ratio of variation between two variances. The "f_classif()" method from the sci-kit-learn library is used to calculate ANOVA F-scores for each feature in relation to the target variable in machine learning. The analysis of variance (ANOVA) F-test can be written as

$$\text{F} = \frac{\text{variance between the groups}}{\text{variance within the groups}} \tag{1}$$

$$\text{variance between the groups} = \frac{\sum_{i=1}^{n} n_i \left(\bar{Y}_i - \bar{Y}\right)^2}{K-1} \tag{2}$$

$$\text{variance within the groups} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_i\right)^2}{N-K} \tag{3}$$

Where $\bar{Y}$ represents the overall mean of the data, N denotes the total sample size, K signifies the number of groups, $\bar{Y}_i$ stands for the mean of the ith group sample, n_i represents the number of observations in the ith group, and Y_{ij} represents the jth observation in the ith group out of K groups [27].

2) Mutual Information (MI)

If two random variables are dependent on each other, MI will find out. Mutual information and the entropy of a random variable are intimately related concepts. To find the relationship between the "features" and "the target feature", the " mutual_info_classif() " method is used to generate the mutual information scores. Equation (4) defines MI between random variables $H(X \mid Y)$ as the product of the entropy H(X) and the conditional entropy. X and Y: "

$$\text{MI}(X : Y) = H(X) - H(X \mid Y) \tag{4}$$

" Equation (5) expresses the entropy $H(X)$ of a discrete random variable X with possible values $\{x_1, x_2, \ldots, x_n\}$:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \tag{5}$$

Equation (6) defines the conditional entropy $H(X \mid Y)$ given Y: "

$$H(X \mid Y) = -\sum_{i=1}^{n} \sum_{j=1}^{n} P(x_i, y_j) \log_2 \left(\frac{P(x_i)}{P(x_i, y_j)} \right) \tag{6}$$

" Here, $P(x_i, y_j)$ denotes the joint probability of $X = x_i$ and $Y = y_j$, and $p(x_i)$ represents the probability mass function of x_i [28].

3) Chi-Square test

The univariate statistical technique is utilized by the Chi-square test for feature selection; this strategy identifies the association between the features. For feature independence determination, chi-square scores are computed using the sci-kit-learn machine library's chi2() function. According to [29], the features with the highest Chi-square scores were chosen using SelectKBest(). "

$$X^2 = \frac{\sum (f_0 - f_E)^2}{f_E} \tag{7}$$

" Here, the observed frequency is denoted as f_0 and the expected frequency is denoted as f_E.

C. *Classification algorithms:*

In our proposed work, we utilized several machine learning algorithms for analysis, including SVM, NB, LR, XGBoost, LGBoost, Adaboost, SGD, and ANN. These algorithms are detailed in the following section.

TABLE I. Heart Disease Dataset Description

Feature	Type	Description	Values
age	Numerical	Age	In days
height	Numerical	Height	In centimeters
weight	Numerical	Weight	In kilograms
gender	Categorical	Gender	Women (1), Men (2)
ap-hi	Numerical	Systolic Blood Pressure	Integer values
ap-lo	Numerical	Diastolic Blood Pressure	Integer values
cholesterol	Categorical	Cholesterol Level	Levels 1, 2, 3
gluc	Categorical	Glucose Level	Levels 1, 2, 3
smoke	Binary	Smoking Habit	Binary value
alco	Binary	Alcohol Consumption	Binary value
active	Binary	Physical Activity	Binary value
cardio	Target	Presence of Disease	1: Disease, 0: No disease

1) Support Vector Machine (SVM)

SVMs distinguish between datasets by extending the use of hyperplanes to nonlinear boundaries. The kernel function is crucial to SVM's efficiency and performance. Picking the correct kernel type is critical for reaching peak performance. Linear, polynomial, and Gaussian kernels were among those utilized in this research [30]. These equations represent the linear, polynomial, and Gaussian kernels, respectively: Equation 8, Equation 9, and Equation 10. Imagine the kernel equations are expressed as, with x_i and x_j as the variables.

$$Linear\ kernel : K(x_i, x_j) = x_i \cdot x_j \,. \tag{8}$$

$$Polynomial\ kernel : K(x_i, x_j) = (x_i \cdot x_j + 1)^d \,. \tag{9}$$

$$Gaussian\ kernel : K(x_i, x_j) = exp(-\gamma \parallel x_i - x_j \parallel^2) \,. \tag{10}$$

"d" is the degree of the polynomial.

2) Naive Bayes

The Naive Bayes classifier predicts outcomes by considering conditional probabilities. It calculates the posterior probability using the prior probability, predictor prior probability, and likelihood [31]. Bayes' theorem, which underpins this process, is formulated as:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \tag{11}$$

3) Logistic regression

When building the logistic regression equation, the maximum likelihood ratio evaluates the statistical importance of variables. The conditional probability $P(Y = 1 \mid X)$, where $X = (X_1, X_2, X_3, \ldots, X_N)$ represents the n risk variables linked to the disease, is computed when [32]. Two possible formulations of the logistic regression model are

$$\log\left(\frac{P(Y = 1 \mid X)}{1 - P(Y = 1 \mid X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_N X_N \tag{12}$$

4) EXtreme gradient boosting tree (XGB)

With XGBoost, you get Extreme gradient boosting is a use case for decision-making trees that have been gradient-boosted. Many supervised learning applications, including regression, classification, and ranking, make extensive use of this popular ML method. With this, we hope to test the computational boundaries of machines that adhere to the most stringent requirements of the gradient boosting system. According to [33], [34], this algorithm generates decision trees in a sequential fashion. Our proposed model use the HyperOpt approach to tune the XGB Classifier's hyperparameters. In this approach, the general tree ensemble model is represented by Eq.13.

$$b_i = \phi(a_i) = \sum_{s=1}^{s} m_s(a_i), m_s \epsilon M \,. \tag{13}$$

5) LGBoost

To improve model efficiency and decrease memory usage, the Light Gradient Boosting Machine (LightGBM) technique uses decision trees. For best performance, LGBM uses a leaf-wise splitting strategy, as opposed to other boosting methods that split trees level-wise or depth-wise. Two new methods, "Exclusive Feature Bundling" and "Gradient-based One-Side Sampling", are presented, which address the drawbacks of histogram-based algorithms [35]. The HyperOpt method is used to tune the hyperparameters of this classifier.

6) AdaBoost

Adaptive boosting, or AdaBoost for short, is a boosting approach used as an ensemble method in machine learning. AdaBoost is an approach to supervised learning that uses sequential growth to address variance and bias. In order to make a more accurate final prediction, this method merges numerous weak classifiers into one "strong" classifier [36].

Here is the final AdaBoost equation:

$$Za = \text{sign}\left(\sum_{p=1}^{P} \Omega_p z_p(a)\right) \tag{14}$$

D. *Stochastic Gradient Descent*

SGD is employed to determine the optimal parameter configuration of a machine learning algorithm. It iteratively adjusts the configuration of the machine learning model to minimize the error rate. SGD operates by updating the network configuration after processing each training point, aiming to locate the global minimum. This approach reduces error by approximating the gradient based on a randomly selected batch, thereby optimizing the model without evaluating the entire dataset each time [37].

1) *Artificial Neural Network (ANN)*

Current neurobiological research provides the basis for ANNs, a computational model that aims to imitate the human brain. The difference between the actual and expected output of each neuron can be calculated by an ANN using training and learning approaches. According to [38], [39], [40], in order to decrease error, the weight of each link is modified starting with the output layer, moving through the hidden layer, and eventually ending up in the input layer. Thus, the accuracy of input pattern identification is enhanced, allowing for the prediction of its probability. The architecture utilized for artificial neural networks (ANN) in this paper is as follows:

TABLE II. Architecture of the Proposed Stroke Prediction Neural Network

Layer (Type)	**Output Dimensions**	**No. of Parameters**
dense_1 (Dense)	(None, 16)	320
dense_2 (Dense)	(None, 32)	544
dropout_1 (Dropout)	(None, 32)	0
batch_norm_1 (Batch Normalization)	(None, 32)	128
dense_3 (Dense)	(None, 64)	2112
dropout_2 (Dropout)	(None, 64)	0
batch_norm_2 (Batch Normalization)	(None, 64)	256
dense_4 (Dense)	(None, 64)	4160
dropout_3 (Dropout)	(None, 64)	0
batch_norm_3 (Batch Normalization)	(None, 64)	256
dense_5 (Dense)	(None, 64)	4160
dropout_4 (Dropout)	(None, 64)	0
batch_norm_4 (Batch Normalization)	(None, 64)	256
dense_6 (Dense)	(None, 64)	4160
dropout_5 (Dropout)	(None, 64)	0
batch_norm_5 (Batch Normalization)	(None, 64)	256
dense_7 (Dense)	(None, 1)	65
Total Parameters		18,585
Trainable Parameters		18,123
Non-trainable Parameters		462

The following are the parameters: Adam is the optimizer, the learning rate is 0.001, the activation function is Relu, the loss function is binary_crossentropy, and the number of epochs is 100.

4. Proposed Model of Heart Disease Prediction with Hyper-parameter tuning using Hybrid Accumulated Feature Selection (HAFS) Meth od

This section describes the proposed model's workflow, dataset description, and preprocessing. The **??**, the process consists of preprocessing, the HAFS framework for feature selection, and the application of ML algorithms.

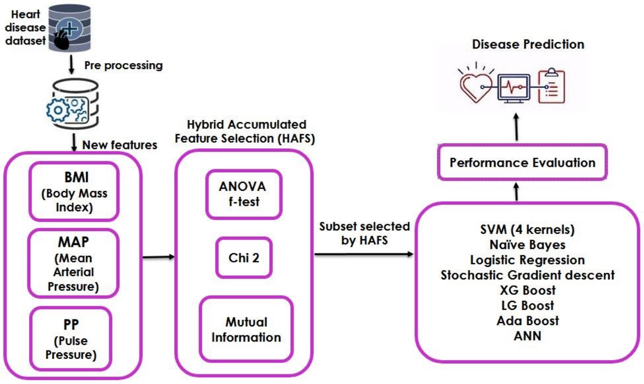

Figure 1. Data Analysis using BMI

A. *New feature creation*

Through feature engineering techniques, we have introduced new features aimed at enhancing the accuracy of our models. BMI serves as an indicator of body fat percentage and correlates with the likelihood of developing associated diseases. Individuals with higher BMI values often face elevated risks, particularly concerning heart disease [41]. We have added BMI and MAP to the dataset. MAP measurement provides insights into blood flow dynamics, vascular resistance, and arterial pressure levels. It holds critical importance as a therapeutic target for “heart diseases”. MAP values are typically calculated using “systolic and diastolic blood pressure” readings [42]. The measurement of PP may indicate the likelihood of developing heart disease. A person’s pulse pressure is currently an indication of their risk of coronary heart disease, especially when they are middle-aged or older [43]. This is shown in Algorithm 1. Formulas for BMI, MAP, and PP are written as

$$\text{BMI} = \frac{\text{weight}}{\text{height}^2} \tag{15}$$

$$MAP = \frac{systolic\_pressure + 2 \times diastolic\_pressure}{3} \tag{16}$$

$$PP = systolic - diastolic \tag{17}$$

Algorithm 1 Feature Engineering Algorithm for Cardiovascular Disease Prediction

```
Input  : Dataset $D$ with features (height, weight, systolic
         pressure, diastolic pressure)
Output: Enhanced dataset $D'$ with new features (BMI,
         MAP, PP)
while enhancement not complete do
    Feature Engineering Steps:
        1) Calculate BMI from height and weight
           $\text{BMI} = \frac{\text{weight}}{\text{height}^2}$
        2) Add BMI to the dataset $D'$
        3) Calculate MAP from systolic and diastolic pressure
           $\text{MAP} = \frac{\text{systolic pressure}+2\times\text{diastolic pressure}}{3}$
        4) Add MAP to the dataset $D'$
        5) Calculate Pulse Pressure (PP) from systolic and
           diastolic pressure
           PP = systolic pressure – diastolic pressure
        6) Add PP to the dataset $D'$
end
```

B. *Hybrid Accumulated Feature Selection (HAFS) framework*

With feature selection algorithms, the most prominent features are selected to increase classification accuracy and reduce classification time. The proposed HAFS framework uses three feature ranking methods, such as the ANOVA f-test, Mutual Information (MI), and Chi-Squared test. According to the results, it appears that feature ranks are quite similar for all methods. As the individual feature selection methods are producing similar results, the proposed HAFS methodology combines their ranks and takes the cumulative rank to select the top features. The ranks of features are represented in Table III. Based on the ranks, the top 10 features are considered the optimal features and selected for improving classification accuracy. This is shown in Algorithm 2. The results for the HAFS framework are explained in detail in the result section.

TABLE III. Hybrid Accumulated Feature Selection (HAFS) Method for Feature Selection

Features	ANOVA f-test ranks	Mutual Information ranks	Chi 2-test ranks	Sum
Systolic	1	1	1	3
MAP	2	2	3	7
Age	5	5	6	16
Pulse Pressure	4	4	2	10
Diastolic	3	3	4	10
BMI	7	7	7	21
Weight	8	8	5	21
Cholesterol	6	6	8	20
Glucose	9	9	9	27
Active	10	12	10	32
Height	11	13	12	36
Smoke	12	11	11	34
Gender	13	10	14	37
Alcohol	14	14	13	41

```
Algorithm 2 HAFS (Hybrid Accumulated Feature Selection) Framework
Input  : Dataset $D$ with features
Output: Optimal feature subset $F_{opt}$
Feature Selection Steps:
  1) Perform ANOVA f-test, MI, and Chi-Squared test on features in $D$
     - Obtain ranks for each feature based on these tests: $R_f^{ANOVA}$, $R_f^{MI}$, $R_f^{Chi^2}$
  2) Calculate cumulative rank for each feature: $R_f = R_f^{ANOVA} + R_f^{MI} + R_f^{Chi^2}$
  3) Select top features based on cumulative ranks
     - Choose the top 10 features with the lowest cumulative ranks: $F_{opt} = \{f | R_f \text{ is minimal for } f\}$
  4) Formulate optimal feature subset $F_{opt}$
     - Include features selected in the previous step: $F_{opt} = \{f_1, f_2, ..., f_{10}\}$
```

C. HyperOpt tunning

Model building begins with feature engineering and is then followed by hyperparameter optimization. ML models without proper hyperparameter tuning have very low chances of getting accurate results [44]. Tuning hyperparameters becomes essential for ML methods since default hyperparameters cannot guarantee performance [45]. Using the HyperOpt algorithm, hyperparameters were tuned for each model. This particular application of HyperOpt will be particularly useful due to its versatility in adapting to a variety of parameters [46], [47]. The main functions in HyperOpt is hp.choice(), hp.radient(), hp.uniform(), and hp.normal(). The results of the hyperOpt model are explained in section V results and analysis.

```
Algorithm 3 Tuned Model
Input  : Dataset $D$ with features
Output: Tuned machine learning model
Algorithm Steps:
  1) Perform feature engineering on dataset $D$ by creating new features
     - Create new features to enhance model performance. For example, calculate Body Mass Index (BMI) as:
       $$\text{BMI} = \frac{\text{weight}}{\text{height}^2}$$
     Incorporate BMI into $D$ to capture body fat indicators.
  2) Conduct hyperparameter optimization for machine learning models
     - Utilize the HyperOpt algorithm to find optimal parameters. Hyperparameters are chosen from distributions such as uniform ($\mathcal{U}(a, b)$), normal ($\mathcal{N}(\mu, \sigma)$), and gradient-based methods and the function can be formulated as:
       $$\theta^* = \arg\max_{\theta} \mathcal{L}(D, \theta)$$
     where $\mathcal{L}$ denotes the model's likelihood or another appropriate objective function.
  3) Apply the tuned hyperparameters to machine learning models
     - Implement the optimized settings in models to maximize performance metrics like accuracy, precision, and recall.
  4) Evaluate the performance of the tuned models
     - Assess model performance using metrics such as accuracy, precision, recall, and F1-score:
       $$\text{F1-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$
     Compare these metrics against baseline results to quantify improvement.
```

5. Experimentation, Performance metrics, and Results

The subsequent part presents details regarding the necessary experimental setup, several metrics employed to assess performance, and the outcomes of the models.

A. *Experimental Setup*

The performance of the ML model is assessed in the experimental setup portion using Python and the Google Collaboratory environment, also known as ”Google Colab”. This research program focuses on developing ML models using high-performance hardware options such as GPUs. It provides a serverless interactive programming environment based on Jupyter Notebooks. Like the other G Suite products, Google Colab is entirely free to use. Data cleaning, classification, segmentation, prediction, and visualization are just a few of the many uses for the system’s various models. An x64-based processor, 16 GB of RAM, a 64-bit OS, and an “ Intel(R) Core(TM) i5-8250U” central processing unit (CPU) with a base “ frequency of 1.60 GHz ” and a maximum turbo “frequency of 1.80 GHz” were the requirements for the experiment.

B. *Performance Metrics*

The model’s accuracy is evaluated using many measures, most of which are solely determined by the values in the confusion matrix. During the evaluation of the ML models, multiple performance measurements were acquired. This analysis will examine the vocabulary commonly encountered in the relevant studies. Refer to the analysis conducted [48].

Accuracy: The accuracy is the proportion of right predictions out of the total number of positive and negative classifications, as follows in Eq 18.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} * 100\,. \qquad (18)$$

Precision: Precision measures the proportion of true positive predictions among all positive predicted classes, as follows in Eq 18 the formula for precision is denoted as Eq. 19.

$$Precision = \frac{TP}{TP + FP} * 100\,. \qquad (19)$$

Recall: Recall is a measure that quantifies the proportion of accurately predicted positive observations out of all the observations in the actual class as follows in Eq. 20.

$$Recall = \frac{TP}{TP + FN} * 100\,. \qquad (20)$$

F1-score: F1-score is determined by utilizing recall and precision as follows in Eq. 21.

$$F1 - Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}\,. \qquad (21)$$

C. *Results and Comparative Analysis*

This section provides the data pre-processing and experimental results obtained by applying the HAFS model. In exploratory data analysis, it is evident that the dataset exhibits a balanced distribution. Specifically, there is a slight difference between 49.5% of individuals without heart disease and 50.4% who have been diagnosed with it. There are some discrepancies noticed in the dataset, with heights lower than 125 or greater than 210 for the height attribute and those are removed. Similarly, those regions with a higher diastolic blood pressure have been eliminated, as the systolic blood pressure should be higher. For the remaining records, no missing or null values were found in the dataset. From 70000 original samples, the dataset was reduced to 68413 samples following the data cleaning process. As explained in Section IV we have created three new features such as BMI, MAP, and PP. The BMI analysis represented in Figure 2 that if a person is overweight or obese, they are more likely to have a CVD than a person who has a normal and underweight BMI. The MAP analysis is represented in Figure 3 showing that most of the people with high MAP have more chances to get heart disease than normal people. The PP analysis in Figure 4 shows that people who have high pulse pressure are having greater chance to get heart disease.

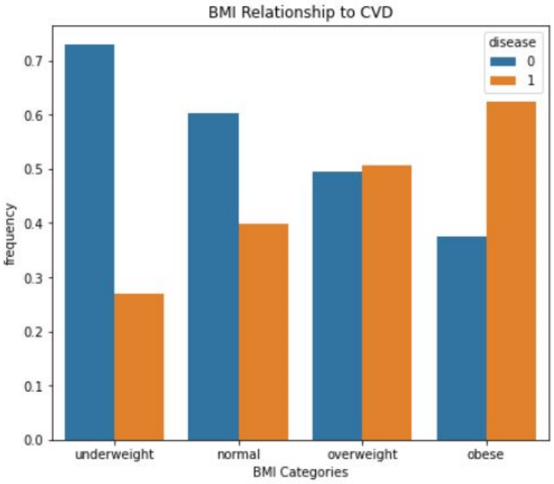

Figure 2. Data Analysis using BMI

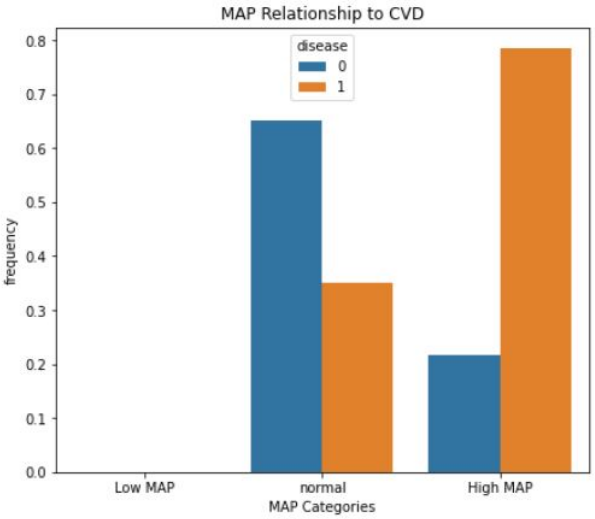

Figure 3. Data Analysis using MAP

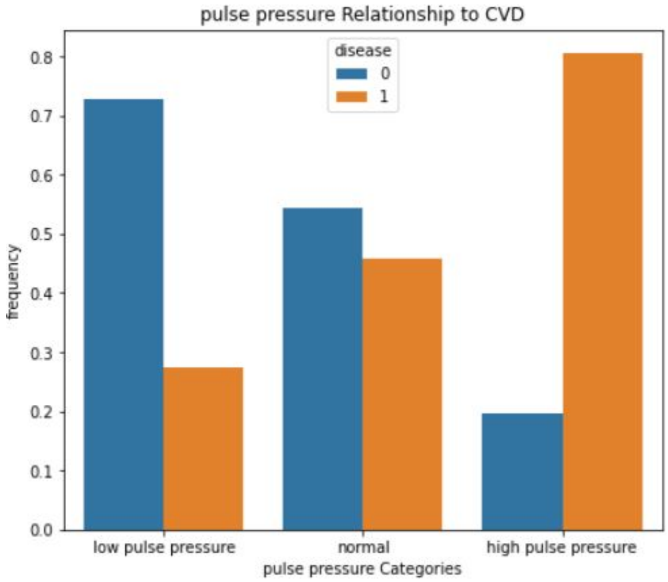

Figure 4. Data Analysis using PP

Feature ranking and selection are done by the HAFS framework as explained in section IV. The top 10 features such as systolic, MAP, age, PP, diastolic, BMI, weight, cholesterol, glucose, and active are selected as optimal features for increasing the performance of a model. As for the remaining four features, they are eliminated since comprehensive details on smoking and alcohol are required. It is also true that eliminating features can have some impact on models' performance. However, we should get complete information, such as how often a person smokes and how many cigarettes they consume daily. When it comes to alcohol consumption, it should be collected like the alcohol intake per day, as how frequently a person consumes alcohol. Figure 5 describes the graphical representation of HAFS framework.

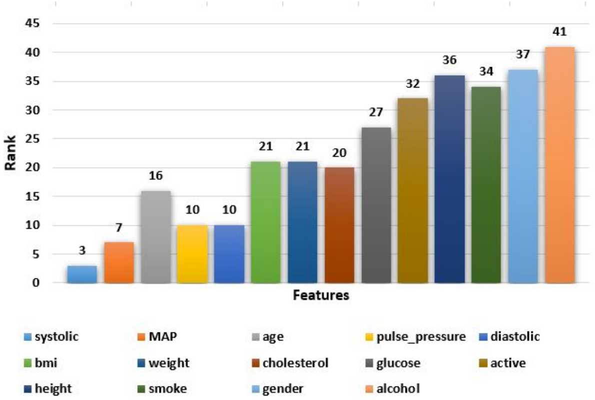

Figure 5. Data Analysis using PP

Following feature selection, the dataset is 80:20 split across the training and testing sets. Eighty percent of the data set is used to train the model, while twenty percent is used to evaluate its performance. To improve the outcomes, 10-fold cross-validation is used. Several algorithms are applied to the reduced feature subset such as SVM, NB, LR, XGBoost, LGBoost, AdaBoost, SGD, and ANN to find the best classification method for classification. Parameter tuning is done by using the HyperOpt technique in this work. Various SVM kernels such as linear, polynomial and Gaussian kernels are applied to the dataset and results showed that the linear kernel achieves the highest accuracy at 73% among the remaining two kernels. The performance analysis of all algorithms used in this model is represented in Table IV.

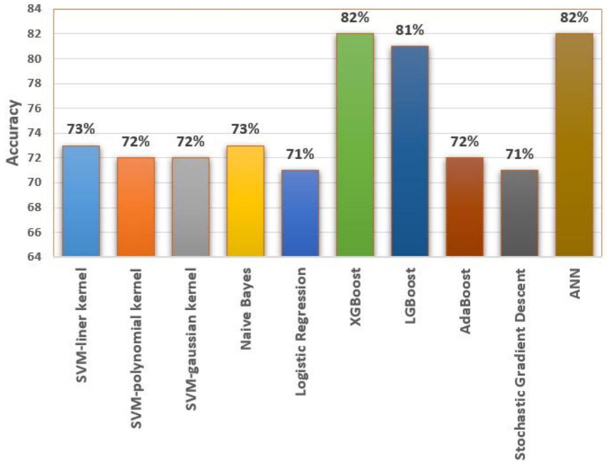

Figure 6. Data Analysis using PP

TABLE IV. Performance Analysis of Various Algorithms

Algorithm name	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVM-linear kernel	73	70	81	75
SVM-polynomial kernel	72	67	86	76
SVM-gaussian kernel	72	69	82	75
Naive Bayes	73	69	83	75
Logistic Regression	71	70	73	71
XGBoost	82	79	86	83
LGBoost	81	78	86	82
AdaBoost	72	69	81	75
Stochastic Gradient Descent	71	67	83	75
ANN	82	79	80	82

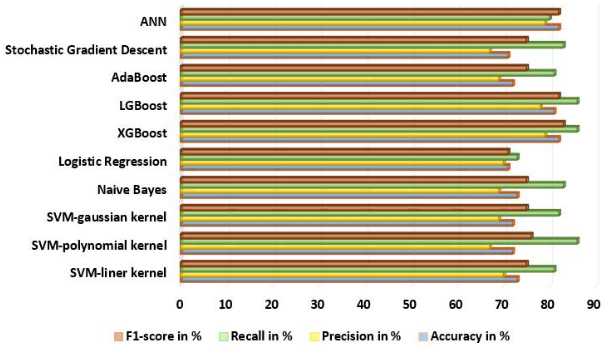

Figure 7. Data Analysis using PP

Figure 6 shows that among all algorithms used in this proposed work, XGBoost with HyperOpt and ANN achieved the highest accuracy at 82% and LGBoost achieved the next highest accuracy at 81%. ANN with two hidden dense layers over 100 epochs achieved the highest accuracy at 82%, precision at 79%, recall at 80%, and f1-score at 82%. Batch normalization is applied to the model for performance improvement. Performance comparison for all algorithms is presented in Figure 7 in a detailed manner. As described in table 3, the highest precision is achieved by XGBoost with HyperOpt parameter tunning and ANN as 79%, the highest recall is achieved by SVM polynomial kernel, XGBoost, and LGBoost as 86%, and highest F1-score is achieved by XGBoost as 83%. The least performance was achieved by LR with 71% accuracy, 70% precision, 73% recall, 71% f1-score. Most of the studies have showed good results on small datasets below 1000 records, but the proposed HAFS model is evaluated using over 70000 records data set. Table V describes that the HAFS model achieved a good score when compared with some literature studies.

6. Conclusions

Timely identification and suitable management of cardiovascular illness are crucial in mitigating fatality rates. This work aims to improve classification models by combining feature engineering and hyperparameter optimization techniques. Significantly, the inclusion of recently introduced metrics such as BMI, MAP, and PP has demonstrated their crucial role in improving the performance model. This model evaluated the accuracy of the HAFS model in predicting cardiovascular disease. The HAFS framework utilizes various feature selection techniques, such as the ANOVA f-test, MI, and Chi-Squared test to determine the most significant features. By employing these techniques in combination, we can provide a strong and reliable process for selecting the most important features. This is essential for improving the performance of the model. The proposed model assessed ML models, such as SVM, NB, LR, XGBoost, LGBoost, AdaBoost, SGD, and ANN. The algorithms' performance was evaluated based on various metrics.

The HyperOpt algorithm was utilized to do hyperparameter optimization, a critical process for enhancing model performance. HyperOpt's versatility in handling diverse parameter distributions renders it a highly effective tool for hyperparameter optimization. HyperOpt optimizes model performance by methodically searching the hyperparameter space and identifying the best-tuned values.

The empirical findings indicated that XGBoost and ANN, with hyperparameters optimized using HyperOpt, surpassed all other models in terms of performance. XGBoost obtained an accuracy of 82%, precision of 79%, recall of 86%, and an F1-score of 83%. In the same manner, the Artificial Neural Network (ANN) attained an accuracy rate of 82%, a precision rate of 79%, a recall rate of 80%, and an F1-score of 82%. LGBoost achieved strong performance, with an accuracy rate of 81%, precision rate of 78%, recall rate of 86%, and an F1-score of 82%. Logistic Regression exhibited the poorest result, achieving an accuracy rate of 71%.

The HAFS model, which includes additional variables such as BMI, MAP, and PP, demonstrated strong performance on an extensive dataset of 70,000 records. This dataset served as a thorough platform for assessing the efficacy of the HAFS framework. The work emphasizes the significance of feature engineering and hyperparameter tun-

TABLE V. Comparison of HAFS Performance with Previous Studies

Methodology	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
XGBoost with wrapper methods [[18]	73.74	76.0	69.0	72.0
Bagged Decision Tree [19]	74.8	76.2	67.4	71.5
Neural network [20]	74.9	76.2	68.2	73.0
Optimized Decision tree [21]	73.14	75.35	69.22	77.63
Neural networks [22]	71.82	72.0	72.0	72.0
Random forest [24]	72.69	74.85	69.48	-
Proposed HAFS model with XGBoost	82	79	86	83
Proposed HAFS model with ANN	82	79	80	82

ing in creating precise predictive models for cardiovascular disease.

Although this work used only one dataset to test the experimental feature selection technique, future validation could include multiple cardiovascular disease datasets of significant size. Conducting comparison assessments would offer more insights into the applicability and strength of the HAFS framework. Further investigation could be conducted in the future to examine the incorporation of additional sophisticated feature selection techniques and optimization algorithms in order to further improve the performance of the model.

To summarize, timely identification and suitable management of cardiovascular illness are crucial in order to decrease mortality rates. This work highlights the crucial importance of feature engineering and hyperparameter adjustment in enhancing the accuracy of predictive models. The combination of the HAFS framework and HyperOpt-tuned models, specifically XGBoost and ANN, exhibited exceptional performance in the classification of cardiovascular illness. These findings support the ongoing endeavours to create dependable and efficient instruments for the early identification and intervention of cardiac disease.

In conclusion, future work should involve applying the HAFS framework to multiple large-scale datasets to validate its generalizability and robustness. Investigating advanced feature selection techniques and hyperparameter optimization methods can further enhance model performance. Developing ensemble models and incorporating deep learning architectures may also improve accuracy. Real-time data integration and clinical trials can assess practical applicability while addressing ethical and privacy concerns to ensure compliance with data protection regulations.

7. Declarations

a) *Confict of Interest*

The authors affirm that there are no conflicts of interest associated with this publication.

b) *Acknowledgments*

We are grateful for the assistance provided by the Indian Institute of Information Technology, Nagpur, Visvesvaraya National Institute of Technology, Nagpur, and Symbiosis Institute of Technology, Nagpur, India.

c) *Data availability statement*

The article contains all pertinent data that substantiates the conclusions of this investigation.

d) *Funding*

Public, commercial, or non-profit funding bodies did not provide any specific funding for this study.

References

[1] L. R. Teixeira *et al.*, "The effect of occupational exposure to noise on ischaemic heart disease, stroke and hypertension: A systematic review and meta-analysis from the who/ilo joint estimates of the work-related burden of disease and injury," *Environment international*, vol. 154, p. 106387, 2021.

[2] C. Guo *et al.*, "Recursion enhanced random forest with an improved linear model (rerf-ilm) for heart disease detection on the internet of medical things platform," *IEEE Access*, vol. 8, pp. 59 247–59 256, 2020.

[3] M. Kivimäki and I. Kawachi, "Work stress as a risk factor for cardiovascular disease," *Current cardiology reports*, vol. 17, no. 9, pp. 1–9, 2015.

[4] J. Fuller, L. Stevens, and S. Wang, "Risk factors for cardiovascular mortality and morbidity: The who multinational study of vascular disease in diabetes," *Diabetologia*, vol. 44, no. 2, pp. S54–S64, 2001.

[5] S. B. Shinde, K. Lahari, K. C. Garimella, V. S. Sree, N. K. Pikle, G. S. Bhavekar, P. Borkar, S. Badhiye, and M. Raghuwanshi, "Experimental analysis of heart disease prediction using machine learning with emphasis on hyper parameter tuning and recursive feature elimination." *International Journal of Intelligent Engineering & Systems*, vol. 16, no. 6, 2023.

[6] S. Shinde, M. P Kurhekar, M. Gulhane, and N. K Pikle, "Design of a novel enhanced machine learning model for early prediction of cerebral stroke," *International Journal of Computing and Digital Systems*, vol. 16, no. 1, pp. 1–22, 2024.

[7] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE access*, vol. 7, pp. 81 542–81 554, 2019.

[8] D. Y. Omkari and S. B. SHINDE, "Opportunities and challenges of machine learning and deep learning techniques in cardiovascular disease prediction: A systematic review," *Journal of Biological Systems*, vol. 31, no. 02, pp. 309–344, 2023.

[9] D. Y. Omkari and S. B. Shinde, “Cardiovascular disease prediction using machine learning techniques with hyperopt,” in *International Conference on Communication and Intelligent Systems*. Springer, 2022, pp. 585–597.

[10] G. S. Bhavekar and A. D. Goswami, “A hybrid model for heart disease prediction using recurrent neural network and long short term memory,” *International Journal of Information Technology*, pp. 1–9, 2022.

[11] D. Peddireddy *et al.*, “Deep learning based approach for identifying conventional machining processes from cad data,” *Procedia Manufacturing*, vol. 48, pp. 915–925, 2020.

[12] S. S. Arief Kanza Rafly, Udin Harun AL Rasyid M, “Efficeint early detection of patient diagnosis and cardiovascular disease using an iot system with machine learning and fuzzy logic,” *International Journal of Computing and Digital Systems*, vol. 16, no. 01, pp. 183–199, 2024.

[13] G. Saranya and A. Pravin, “Hybrid global sensitivity analysis based optimal attribute selection using classification techniques by machine learning algorithm,” *Wireless Personal Communications*, vol. 127, no. 3, pp. 2305–2324, 2022.

[14] R. Das, I. Turkoglu, and A. Sengur, “Effective diagnosis of heart disease through neural networks ensembles,” *Expert systems with applications*, vol. 36, no. 4, pp. 7675–7680, 2009.

[15] L. Ali *et al.*, “An optimized stacked support vector machines based expert system for the effective prediction of heart failure,” *IEEE Access*, vol. 7, pp. 54 007–54 014, 2019.

[16] X. Yuan *et al.*, “A stable ai-based binary and multiple class heart disease prediction model for iomt,” *IEEE Transactions on Industrial Informatics*, vol. 18, no. 3, pp. 2032–2040, 2021.

[17] Z. Arabasadi *et al.*, “Computer aided decision making for heart disease detection using hybrid neural network-genetic algorithm,” *Computer methods and programs in biomedicine*, vol. 141, pp. 19–26, 2017.

[18] J. Maiga and G. G. Hungilo, “Comparison of machine learning models in prediction of cardiovascular disease using health record data,” in *2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*. IEEE, 2019, pp. 54–58.

[19] N. Hasan and Y. Bao, “Comparing different feature selection algorithms for cardiovascular disease prediction,” *Health and Technology*, vol. 11, no. 1, pp. 49–62, 2021.

[20] V. Shorewala, “Early detection of coronary heart disease using ensemble techniques,” *Informatics in Medicine Unlocked*, vol. 26, p. 100655, 2021.

[21] V. Cheekati, D. D. Natarajasivan, and S. Indraneel, “Ensemble approaches can aid in the early detection of coronary heart disease,” *NVEO-NATURAL VOLATILES ESSENTIAL OILS Journal— NVEO*, pp. 12 224–12 239, 2021.

[22] B. Martins *et al.*, “Data mining for cardiovascular disease prediction,” *Journal of Medical Systems*, vol. 45, no. 1, pp. 1–8, 2021.

[23] S. Ouf and A. IB, “A proposed paradigm for intelligent heart disease prediction system using data mining techniques,” *Journal of Southwest Jiaotong University*, vol. 56, no. 4, 2021.

[24] R. Hagan, C. J. Gillan, and F. Mallett, “Comparison of machine learning methods for the classification of cardiovascular disease,” *Informatics in Medicine Unlocked*, vol. 24, p. 100606, 2021.

[25] J. Ali *et al.*, “A comparative study of machine learning algorithms to detect cardiovascular disease with feature selection method,” in *Machine Intelligence and Data Science Applications*. Springer, Singapore, 2022, pp. 573–586.

[26] K. Mahalakshmi and P. Sujatha, “Critical analysis of feature selection methods for data preprocessing with heart disease dataset,” in *Data Intelligence and Cognitive Informatics*. Springer, Singapore, 2022, pp. 667–682.

[27] M. S. Pathan *et al.*, “Analyzing the impact of feature selection on the accuracy of heart disease prediction,” *Healthcare Analytics*, vol. 2, p. 100060, 2022.

[28] M. Kavitha *et al.*, “Hybrid model using feature selection and classifier in big data healthcare analytics,” in *Inventive Communication and Computational Technologies*. Springer, Singapore, 2022, pp. 777–791.

[29] H. A. Mengash *et al.*, “Smart cities-based improving atmospheric particulate matters prediction using chi-square feature selection methods by employing machine learning techniques,” *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2067647, 2022.

[30] K. Harimoorthy and M. Thangavelu, “Multi-disease prediction model using improved svm-radial bias technique in healthcare monitoring system,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 3, pp. 3715–3723, 2021.

[31] K. Velswamy *et al.*, “Classification model for heart disease prediction with feature selection through modified bee algorithm,” *Soft Computing*, vol. 26, no. 23, pp. 13 049–13 057, 2022.

[32] M. S. Amin, Y. K. Chiam, and K. D. Varathan, “Identification of significant features and data mining techniques in predicting heart disease,” *Telematics and Informatics*, vol. 36, pp. 82–93, 2019.

[33] K. Budholiya, S. K. Shrivastava, and V. Sharma, “An optimized xgboost based diagnostic system for effective prediction of heart disease,” *Journal of King Saud University-Computer and Information Sciences*, 2020.

[34] L. Gao and Y. Ding, “Disease prediction via bayesian hyperparameter optimization and ensemble learning,” *BMC research notes*, vol. 13, no. 1, pp. 1–6, 2020.

[35] D. D. Rufo *et al.*, “Diagnosis of diabetes mellitus using gradient boosting machine (lightgbm),” *Diagnostics*, vol. 11, no. 9, p. 1714, 2021.

[36] T. R. Mahesh *et al.*, “Adaboost ensemble methods using k-fold cross validation for survivability with the early detection of heart disease,” *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[37] R. Kannan and V. Vasanthi, “Machine learning algorithms with roc curve for predicting and diagnosing the heart disease,” in *Soft computing and medical bioinformatics*. Springer, 2019, pp. 63–72.

[38] O. W. Samuel *et al.*, “An integrated decision support system based on ann and fuzzy_ahp for heart failure risk prediction,” *Expert Systems with Applications*, vol. 68, pp. 163–172, 2017.

[39] I. D. Mienye, Y. Sun, and Z. Wang, “Improved sparse autoencoder

based artificial neural network approach for prediction of heart disease," *Informatics in Medicine Unlocked*, vol. 18, p. 100307, 2020.

[40] F. Ali *et al.*, "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion," *Information Fusion*, vol. 63, pp. 208–222, 2020.

[41] A. Sommer and G. Twig, "The impact of childhood and adolescent obesity on cardiovascular risk in adulthood: a systematic review," *Current diabetes reports*, vol. 18, no. 10, pp. 1–6, 2018.

[42] D. DeMers and D. Wachs, "Physiology, mean arterial pressure," *StatPearls [Internet]*, 2021.

[43] M. Pareek *et al.*, "Pulse pressure, cardiovascular events, and intensive blood-pressure lowering in the systolic blood pressure intervention trial (sprint)," *The American Journal of Medicine*, vol. 132, no. 6, pp. 733–739, 2019.

[44] P. Srinivas and R. Katarya, "hyoptxg: Optuna hyper-parameter optimization framework for predicting cardiovascular disease using xgboost," *Biomedical Signal Processing and Control*, vol. 73, p. 103456, 2022.

[45] H. Cho *et al.*, "Basic enhancement strategies when using bayesian optimization for hyperparameter tuning of deep neural networks," *IEEE Access*, vol. 8, pp. 52 588–52 608, 2020.

[46] S. Putatunda and K. Rama, "A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of xgboost," in *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning*, 2018, pp. 6–10.

[47] R. Andonie, "Hyperparameter optimization in learning systems," *Journal of Membrane Computing*, vol. 1, no. 4, pp. 279–291, 2019.

[48] R. R. Sanni and H. S. Guruprasad, "Analysis of performance metrics of heart failured patients using python and machine learning algorithms," *Global transitions proceedings*, vol. 2, no. 2, pp. 233–237, 2021.

D.YASO OMKARI D.YASO OMKARI, received the bachelor's degree Master's degree in engineering from Jawaharlal Nehru Technological University-Anantapur. She is currently pursuing the Ph. D degree in computer science with the VIT-AP University Near Vijayawada, Andhra Pradesh. Her current research interests include Data mining, Bigdata, Nature inspired algorithms, Machine learning, Deep Learning, and Optimization Techniques.

Tausif Diwan Dr. Tausif Diwan received his M.Tech. in Computer Science and Ph.D. in Parallel Computing from Department of Computer Science and Engineering, Visvesvaraya National Institute of Technology, Nagpur, India in 2011 and 2017 respectively. Nagpur as an Assistant Professor and currently working as Associate Dean Associate Dean at IIIT Nagpur. His research areas include parallel computing, machine learning, and deep learning. He has published several research papers in various international conferences and reputed journals. An Australian Patent titled "An IoT-based Smart Food-Cold Chain Transportation System" is also granted to his account. He has associations and collaborations with several industries. He is an NVidia-certified DLI instructor for the "Foundation of deep learning" and "Building Transformer-Based Natural Language Processing Applications" courses and he secured the "Platinum" award as a DLI Instructor from Nvidia in the year 2022.

Snehal B Shinde Dr. Snehal B. Shinde is a distinguished academic with a Ph.D. in Computer Science and Engineering from VNIT, Nagpur, focusing on Systems Biology. He has taught at notable institutions such as Vishwakarma University and Vellore Institute of Technology. His research spans machine learning and deep learning in healthcare, with numerous publications in esteemed journals and conferences. Dr. Shinde is also a co-inventor of a patented system for ECG classification using deep learning. Currently, she is working as an Assistant Professor at IIIT Nagpur.

Pradnya Borkar is a prominent academic with a Ph.D in Computer Science and Engineering from RTMNU,Nagpur focusing on Bioinformatics. Her research area focuses on Parallel Computing, High Performance Computing and Bioinformatics with numerous publications in esteemed and conferences. Dr. Borkar is also a co-inventor of a patented system for An artificial intelligence enabled robot for home assistance.Also She has co-authored one book and edited one book.

Nileshchandra Pikle Dr Nileshchandra Pikle is a distinguished figure in the field of computer science, renowned for his expertise in deep learning and GPU computing. Holding a PhD in Computer Science and Engineering, Dr. Pikle specializes in GPU programming, with a particular focus on the parallelization of Finite Element Method (FEM) on CUDA-enabled GPUs, a subject on which he has published papers