

Received 19 March 2022, accepted 30 March 2022, date of publication 8 April 2022, date of current version 4 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3165792

Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV

GHULAB NABI AHMAD¹, HIRA FATIMA¹, SHAFIULLAH²,
ABDELAZIZ SALAH SAIDI^{3,4}, (Member, IEEE),
AND IMDADULLAH⁵, (Senior Member, IEEE)

¹Institute of Applied Sciences, Mangalayatan University, Aligarh, Uttar Pradesh 202146, India

²Department of Mathematics, K.C.T.C College, Raxaul, A Constituent Unit of BRA, Bihar University, Muzaffarpur, Bihar 842001, India

³Department of Electrical Engineering, King Khalid University, Abha 61411, Saudi Arabia

⁴Laboratoire des Systèmes Electriques, Ecole Nationale d'Ingénieurs de Tunis, Université de Tunis El Manar, Tunis 1068, Tunisia

⁵Faculty of Engineering & Technology, Electrical Engineering Section, University Polytechnic, Aligarh Muslim University, Aligarh 202002, India

Corresponding author: Shafiullah (shafi.stats@gmail.com)

This work was supported by the Deanship of Scientific Research, King Khalid University through Research Groups Program under Grant RGP.2/81/43.

ABSTRACT Predicting cardiac disease is considered one of the most challenging tasks in the medical field. It takes a lot of time and effort to figure out what's causing this, especially for doctors and other medical experts. In this paper, various Machine Learning algorithms such as LR, KNN, SVM, and GBC, together with the GridSearchCV, predict cardiac disease. The system uses a 5-fold cross-validation technique for verification. A comparative study is given for these four methodologies. The Datasets for both Cleveland, Hungary, Switzerland, and Long Beach V and UCI Kaggle are used to analyze the models' performance. It is found in the analysis that the Extreme Gradient Boosting Classifier with GridSearchCV gives the highest and nearly comparable testing and training accuracies as 100% and 99.03% for both the datasets (Hungary, Switzerland & Long Beach V and UCI Kaggle). Moreover, it is found in the analysis that XGBoost Classifier without GridSearchCV gives the highest and nearly comparable testing and training accuracies as 98.05% and 100% for both the datasets (Hungary, Switzerland & Long Beach V and UCI Kaggle). Furthermore, the analytical results of the proposed technique are compared with previous heart disease prediction studies. It is evident that amongst the proposed approach, the Extreme Gradient Boosting Classifier with GridSearchCV is producing the best hyperparameter for testing accuracy. The primary aim of this paper is to develop a unique model-creation technique for solving real-world problems.

INDEX TERMS Heart disease, support vector machine (SVM), logistic regression (LR), gradient boosting classifier (GBC), GridSearchCV.

I. INTRODUCTION

The availability of various medical data prompts us to consider whether there are any efficient and effective methods for analyzing this data and deriving potentially innovative and applied knowledge. One of the most significant difficulties for data analytics is the diagnosis of various diseases. The researchers concentrate their efforts in numerous ways, such as developing high-accuracy prediction models, extracting if-then rules, and experimenting with new cut-off values

The associate editor coordinating the review of this manuscript and approving it for publication was Wentao Fan¹.

for key input variables [1]. All directions are essential and can assist in the case of medical test diagnostic effectiveness. According to many medical data released by competent authorities worldwide, the most common cause of heart failure is myocardial infarction. Cardiovascular disease (CVD) is becoming increasingly common [2]. As reported by the WHO, CVDs affected roughly 1.79 crore individuals in 2016, accounting for 31% of all worldwide mortality. Only heart attacks and strokes are responsible for 85% of these deaths. With early diagnosis and good medical care, many lives may be spared. One of the most significant elements is blood pressure [3]. Heart disease can strike at any time without

generating discomfort, and doctors are unable to forecast these silent assaults [4]. By 2030, the number of deaths due to cardiovascular disease is expected to climb to 23.3 million [5]. The blood veins of the heart carry oxygen, and when these channels become blocked or restricted, heart disease or stroke can occur [6]. High blood pressure, high cholesterol, stress, tension, alcohol usage, sedentary lifestyle, obesity, and diabetes are all factors that affect the heart. These factors help in cardiovascular disease diagnosis. The walls of the arteries thicken when blood pressure rises, causing obstruction and perhaps increasing mortality [7]. Every human being should receive correct treatment before their condition progresses; if a technology exists that can anticipate cardiac disease at an early stage; it might save a million lives. Heart-related disorders can be reduced with the use of medical data and machine learning techniques. Machine learning aids in the finding of patterns in enormous volumes of data that people are unable to detect the use of sophisticated technology aids in the improvement of the current health-care system [8]. A machine learning model learns from the historical data it receives and then creates prediction algorithms to forecast the outcome for fresh data that enters the system as input. The accuracy of these models is determined on the quality and amount of input data. A vast amount of data will aid in the development of a better model that precisely predicts the output [9]. A range of supervised and unsupervised machine learning methods have been utilized by medical academics to identify and predict heart diseases. [10]. The major goal of the various academics' suggested algorithms is to extract patterns from data and forecast future outcomes. Early detection and treatment of cardiac disease can help patients avoid death or lower their mortality rate. Angiography is one of the most popular procedures for diagnosing aberrant narrowing of a cardiac artery. SMO, Nave Bayes, and Ensemble methods were used to evaluate the symptoms, examination, and ECG characteristics, with an accuracy of 88.5 percent in predicting the existence of CAD [11]. Artificial intelligence is being used by governments and corporate health organizations for a variety of purposes, including lowering operating costs and increasing patient satisfaction. Lowering the length of stay and determining the etiology of sickness, among other things to predict early heart attacks, many data mining classification methods were utilized [12]. Finding an ideal solution is quite difficult without thorough data analysis. The purpose of this study is to get as close to zero prediction error rates as feasible. Many academics have previously utilized the dataset that is being examined for this project, which is accessible in the UC Irvine machine learning repository. There are 73 distinct characteristics in the dataset. Machine learning engineers grouped the most optimal qualities from all relevant attributes and created a 14-attribute dataset [13]. Data mining, machine learning, deep learning, and other automated methods for diagnosing cardiac disease are already available. As a result, we'll present a short review of machine learning methodologies in this post. To train the datasets, we leverage machine learning resources. The development

of cardiac disease can be predicted by a number of risk factors. Risk factors include age, sex, blood pressure, cholesterol level, family history of coronary artery disease, diabetes, smoking, alcohol, obesity, heart rate, and chest discomfort etc.

Currently, information about patients with medical reports is widely available in databases in the healthcare sector, and it is rising fast day by day. This unbalanced raw data is significantly redundant. It necessitates pre-processing in order to extract important features, reduce training algorithm execution time, and increase classification efficiency [14]. The current advances in processing capacity and machine learning reprogramming skills improve these processes and open up new research prospects in the healthcare sector [15] particularly in terms of disease early detection, such as cardiovascular disease and cancer, in order to enhance survival rates. Machine learning is utilised in a wide range of applications, from identifying disease risk factors to developing superior automotive safety systems. To solve the current restrictions, machine learning provides the most used prediction modelling tools [16]. It has a lot of potential for transforming massive data and developing prediction algorithms. It uses a computer to learn complicated and non-linear relationships between attributes by minimising the difference between projected and observed results [17]. The machine learns patterns from current dataset features and applies them to an unknown dataset to predict the outcome. One of the most effective machine learning prediction algorithms is classification. When taught with appropriate data, classification is a supervised machine learning method that is excellent at identifying the diseases [18].

The main contribution of this research was to use modern machine learning techniques to construct an intuitive medical prediction system for the diagnosis of heart disease. Different types of machine learning classifier algorithms were trained in this study, including logistic regression (LR), K-nearest neighbours (K-NN), support vector machine (SVM), and Gradient Boosting Classifier (GBC) with and without GridSearchCV, to select the best predictive model for accurate heart disease detection at an early stage.

To achieve the ideal collection of attributes that strongly influenced the performance of the classifiers when predicting the target class, four model selection strategies were used, including correlation-based feature subset evaluator and Gradient Boosting Classifier evaluator. Finally, the whole attribute set and optimal sets obtained via attribute evaluators were used to tune the hyperparameter "GridSearchCV" in the GBC classifier.

One of the most challenging issues in medicine is predicting cardiac disease. It takes a lot of time and effort to figure out what's causing this, especially for doctors and other medical professionals. Researchers used a range of algorithms, including LR, KNN, SVM, and GBC, as well as the GridSearchCV, to predict cardiac disease.

The following is a breakdown of how the paper is structured: Section II examines previous heart disease research

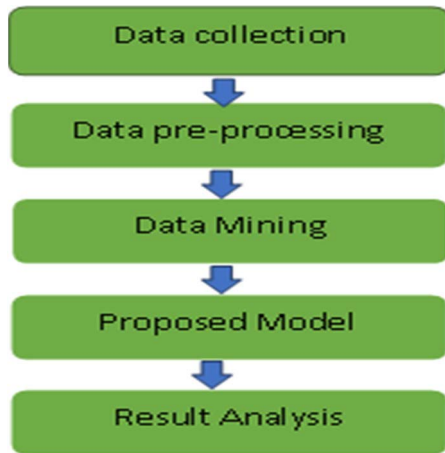


FIGURE 1. The suggested model's flow diagram.

using a variety of machine learning approaches, Section III explains the database we use and our recommended model's approach analysis, and Section IV concludes with detailed results and comparisons to other methods. Finally, in Section V, the paper's conclusion and future study potential are discussed.

II. REVIEW LITERATURE

In the medical area, a lot of works have been done on diseases prediction systems employing various machine learning techniques. Vikas Chourasia *et al.* [19] presented another work on Early Prediction of Heart Diseases Using Data Mining Techniques. They used classification and regression trees, iterative dichotomization, and decision tables, among other approaches. The approach uses 10-fold cross validations, with CART achieving the highest accuracy of 83 percent. Alizadeh Sani *et al.* [20] employed a rule-based classifier, a cost-sensitive method, and Sequential minimum optimization to identify coronary artery disease (SMO), Lakshmana Rao and others [21]. Machine Learning Techniques for Heart Disease Identifying the elements that are most likely to cause heart disease (circulatory strain, diabetes, current smoker, high cholesterol, etc.). As a result, distinguishing between heart disease and other conditions is challenging. The severity of cardiac disease in patients has been determined using a range of data mining and neural networks. CHD sickness is a difficult idea, and as a result, it should be addressed with caution. Failure to recognize the disease early on might have serious consequences for the heart or even result in death. Data burrowing is also used in pharmaceutical research to locate various sorts of metabolic machine learning, which is a strategy that allows a framework to learn from prior data tests and models without having to be actively altered. Logic becomes reliant on machine learning. To construct an HD prediction system, Mohan *et al.* [22] created a hybrid machine learning technique. He also demonstrated a unique technique for collecting essential characteristics from data

in order to train and evaluate machine learning classifiers. They were properly categorized 88.07% of the time. Olaniyan *et al.* [23] developed a three-phase strategy for HD prediction in angina that achieved 88.89 % accuracy using an artificial neural network technology. Ratnasari *et al.* [24] utilised a Gray-level threshold of 150, which was based on PCA and ROI, to decrease X-ray image features. A 13-feature dataset was used in the bulk of prior research. Every study employs the e classification to determine if a patient has heart disease or not, and one of the most common trends is that Cleveland is the most commonly used dataset. Jansosi *et al.* [25]. UCI machine learning repository-heart disease data set. School, G'arate-Escamila *et al.* [26] with the X2 statistical model, DNN and ANN were applied. The clinical data parameters were utilized to ensure that the forecasts were accurate, Andres *et al.* [27] Harvard Medical School is a prestigious medical school in Boston. Using Hungarian-Cleveland datasets, several machine learning classifiers were used to predict heart disease, and PCA was used for dimensionality reduction and feature selection. For feature extraction, Zhang *et al.* [28] used an AdaBoost classifier with a PCA combination, which enhanced prediction accuracy.

III. MATERIALS AND METHODS

The proposed methodology, which includes dataset definition, data pre-processing, machine learning classifiers, attribute evaluators, and performance measures, is discussed in this part.

A. PROPOSED METHODOLOGY

The main contribution of this research was to use modern machine learning techniques to construct an intuitive medical prediction system for the diagnosis of heart disease. Different types of machine learning classifier algorithms were trained in this study, including logistic regression (LR), K-nearest neighbours (K-NN), support vector machine (SVM), and Gradient Boosting Classifier (GBC) with and without GridSearchCV, to select the best predictive model for accurate heart disease detection at an early stage.

To achieve the ideal collection of attributes that strongly influenced the performance of the classifiers when predicting the target class, four model selection strategies were used, including correlation-based feature subset evaluator and Gradient Boosting Classifier with GridSearchCV. Finally, the whole attribute set and optimal sets obtained via attribute evaluators were used to tune the hyperparameter "GridSearchCV" in the GBC classifier. The system uses a 5-fold cross-validation technique for verification. In the comparison research, these four methodologies were applied. The models' performance is evaluated using the Datasets for Cleveland, Hungary, Switzerland, and Long Beach V, as well as the Dataset Heart Disease UCI Kaggle Dataset. The Extreme Gradient Boosting Classifier with GridSearchCV generated the highest and nearly comparable accuracy results for both the hungary, switzerland & Long Beach V and Heart Disease UCI Kaggle Datasets (100% and 99.03%, respectively).

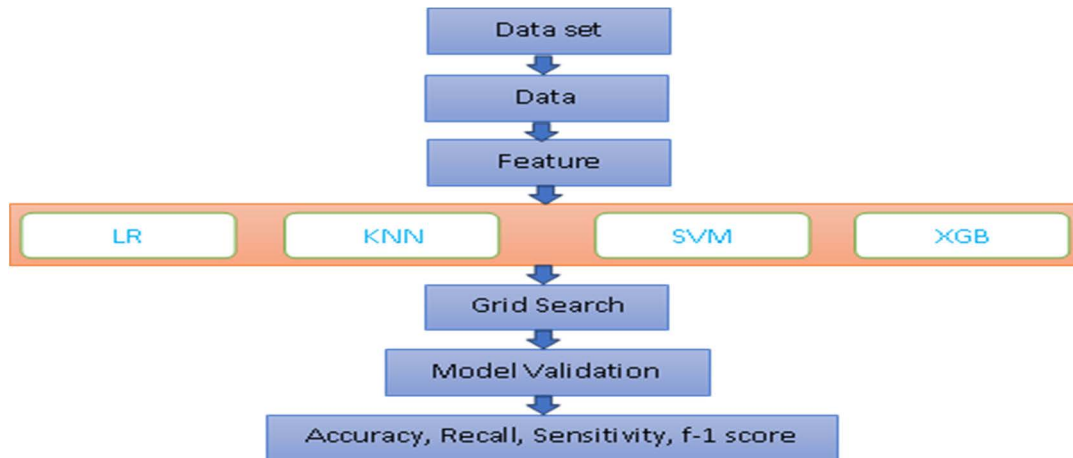


FIGURE 2. The heart disease prediction system proposed model.

The optimal hyperparameter for testing accuracy is the Extreme Gradient Boosting Classifier with GridSearchCV. Many performance indicators, like as Accuracy, Recall, Precision, and F1-Score, may be used to evaluate the models' performance. Previous cardiac prediction studies were compared to the findings. We'd want to develop the model in the future so that it can be used with a variety of feature selection techniques; another option is to utilize a GridSearchCV in conjunction with an Extreme Gradient Boosting Classifier. The study's main purpose is to build on past work by inventing a new and distinctive model-creation approach, as well as to make the model relevant and easy to apply in real-world circumstances. For Cleveland, Hungary, Switzerland, and Long Beach V, the models' performance is evaluated using the GridSearchCV technique with 5-fold cross validation. Our findings imply that cardiac disease may be consistently predicted using four machine learning models and optimization methodologies.

B. METHODOLOGY RECOMMENDED

Creative datasets were acquired, and then the collected data was pre-processed. The sequential forward selection approach was used to pick relevant characteristics. The hyperparameter optimization approaches Grid search were used to optimize the parameters of Logistic Regression, k-Nearest Neighbors, Support Vector Machine, and XGBoost. Finally, the models were tested and analyzed to see if they might predict cardiac disease. The data in our proposed model is validated using 5-fold cross validation. The suggested model's flowchart is shown in **Fig. 2**.

C. DATASET DESCRIPTION

1) CLEVELAND_HUNGARY_SWITZERLAND & LONG BEACH V HEART DISEASE DATASET AND HEART DISEASE UCI KAGGLE DATASET

The dataset utilized in this paper comes from [29], which is a shortened version of [30]. The dataset contains 76

characteristics, including the class attribute, for 1025 patients from Cleveland, Hungary, Switzerland, and Long Beach V, however only a subset of 14 are used in this study. The heart disease prediction dataset has 14 columns, with 13 independent components and one dependent target variable. The target variable is separated into two groups: those who have heart diseases and those who do not. There are 1025 rows in all for 918 patients from heart disease UCI Kaggle dataset, however only a subset of 12 are used in this study. The heart disease prediction dataset has 12 columns, with 11 independent components and one dependent target variable. The target variable is separated into two groups: those who have heart disease and those who do not. There are 918 rows in all. There are no missing values in the dataset. **Table 1** contains a description of the dataset.

2) EXPLORATORY DATA ANALYSIS

To comprehend data insights, some type of data analytics is required, which allows for a better understanding of data patterns. The data distribution of the dataset is examined in the next section. The Cleveland Hungary Switzerland and Long Beach V heart disease dataset comprises 1025 patient records from 713 men and 312 women of various ages, with 499 patients (48.68%) having normal cardiac function and 526 patients (51.32%) suffering cardiac diseases. Among the heart disease patients, there are 300 male patients (57.03%) and 226 female patients (42.97%). The UCI Kaggle dataset on heart disease has 918 patient records from 725 men and 193 women of all ages, with 410 patients (44.67%) having normal cardiac function and 508 patients (55.34 %) having heart disease. **Fig. 3** shows the gender wise distribution, **Fig. 4** the male and female gender distribution (1 = male, 0 = female), **Fig. 5** shows the age distribution of patients with +ve heart diagnosis, **Fig. 6** shows how each feature is connected to the others and aids in determining which attributes are the most essential in this dataset. The values that are created here are less reliant on one another.

TABLE 1. Heart disease dataset attributes and information.

Sr. no.	Feature	Characteristic representation	Specifics
1	Age	Age	Patients age, in years
2	Sex	Sex	0=female; 1=male
3	Chest pain	Cp	4 types of chest pain (1—typical angina; 2— atypical angina; 3—non-anginal pain; 4— asymptomatic)
4	Rest blood pressure	trestbps	Blood pressure normal (in mm Hg on admission to the hospital)
5	Serum cholesterol	chol	Cholesterol level in mg/dl.
6	Fasting blood sugar	fbs	Fasting blood sugar>120 mg/dl (0—false; 1—true)
7	Rest electrocardiograph	restecg	0—normal; 1—having ST-T wave abnormality; 2—left ventricular hypertrophy
8	MaxHeart rate	Thalach	heart rate achieved which is at its maximum value.
9	Exercise-induced angina	exang	Exercise-induced angina (0—no; 1—yes)
10	ST depression	oldpeak	ST depression induced by exercise relative to rest
11	Slope	slope	slope of the peak exercise ST segment (1— upsloping; 2—flat; 3—down sloping)
12	No. of vessels	ca	No. of major vessels (0–3) coloured by fluoroscopy
13	Thalassemia	thal	3 = normal; 6 = fixed defect; 7 = reversable defect.
14	Num(class attribute)	Class	diagnosis of heart disease status (0—nil risk; 1—low risk; 2—potential risk; 3— high risk; 4—very high risk)

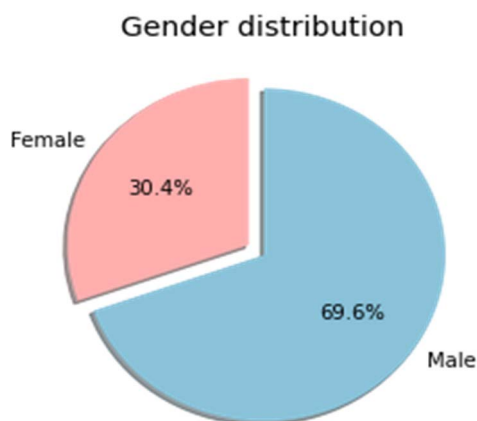


FIGURE 3. Gender wise distribution.

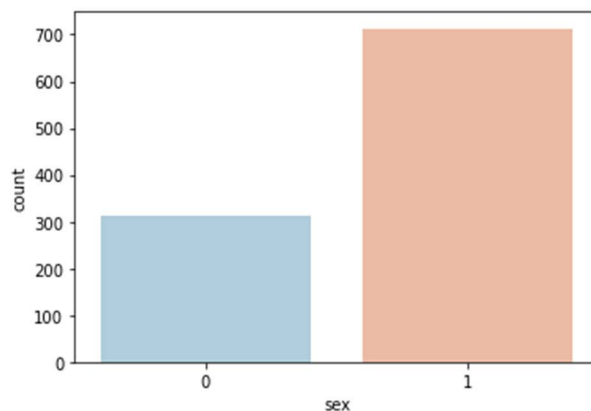


FIGURE 4. The male and female gender distribution (1 = male, 0 = female).

3) METHOD-I LOGISTIC RIGRESSION ANALYSIS

Linear regression is unbounded and the classifier may get error, for this reason logistic regression is used. The most

straight-forward interpretation of logistic regression which resides between 0 and 1 is given in [32], [33]. The model predicts the probability of our random variable that is our target

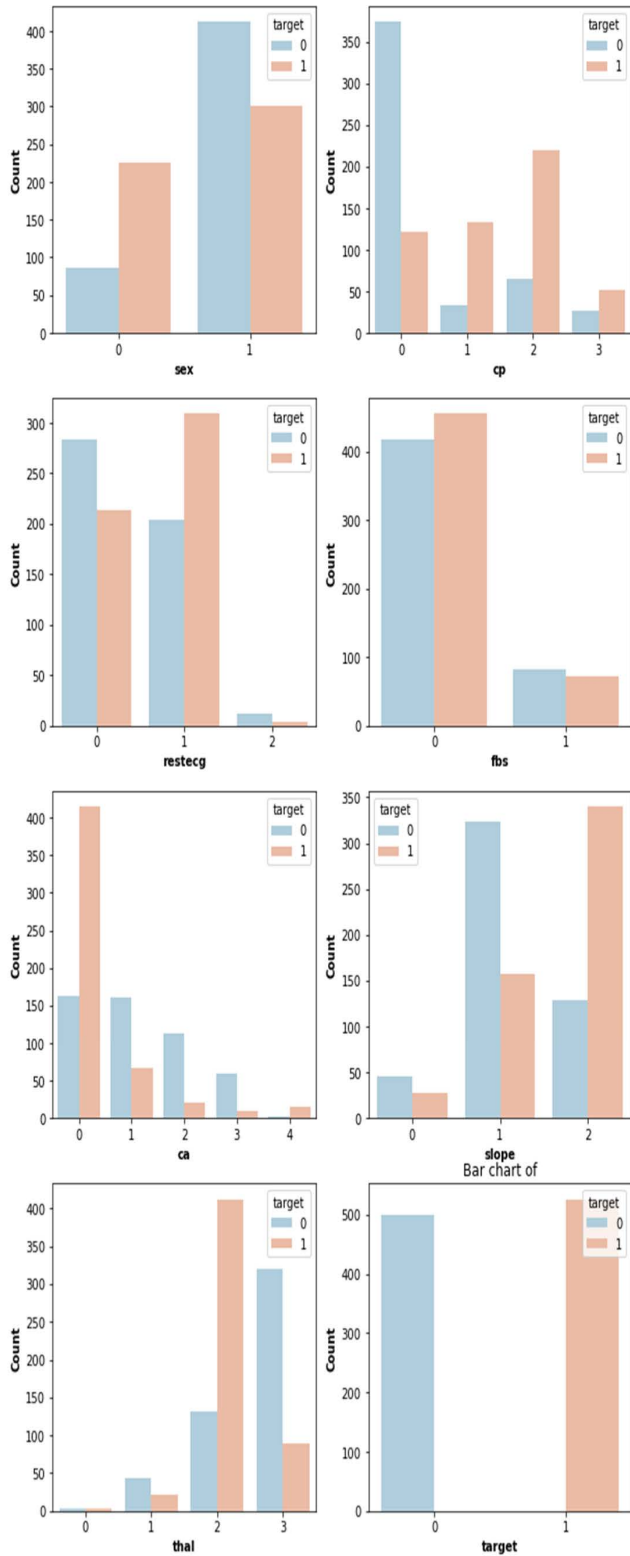


FIGURE 5. Age distribution of patients with +ve heart diagnosis.

value from the given data [34]. The cost function used in logistic regression is sigmoid function. In logistic regression, the linear function is essentially utilized as an input to another

function, such as f in the equation

$$\hat{y} = h_{\theta}(x) = f(\theta^T x) \quad \text{where } 0 \leq h_{\theta}(x) \leq 1 \quad (1)$$

where \hat{y} = predicted value

$$X^T = [1, x_1, x_2, x_3, \dots, x_n],$$

$$\theta^T = [\beta_0, \beta_1, \beta_2, \beta_3, \dots, \beta_n]$$

Then $\Rightarrow \hat{y} = \theta^T X$

$$\begin{aligned} \hat{y} &= \beta_0 + \beta_1 x_1 + \beta_1 x_1 + \beta_1 x_1 + \dots + \beta_n x_n \\ &= \sum_{i=0}^n \beta_n x_n = Z \end{aligned} \quad (2)$$

x = independent variable and θ are coefficient to be learnt [35]. But the linear regression stuff so we need to find a way to cast the logistic regression problem in a manner where by at least the expression above can be used thus if we computed the odds of the outcome as:

$$Odds = \frac{P}{1-P}, \quad 0 \leq p \leq 1 \Rightarrow P = \frac{Odds}{1+Odds} \quad (3)$$

We can move a step closer to casting the problem in a continuous linear manner but this is still just having positive values we need a range of $(-\infty, +\infty)$. That can be done by getting the (natural) logarithm of the odds as:

$$\begin{aligned} \text{logit}(P) &= \log \frac{P}{1-P} = \hat{y} = \theta^T X \\ \log \frac{P}{1-P} &= \theta^T X \Rightarrow \frac{P}{1-P} = e^{\theta^T X} \\ \Rightarrow P &= (1-P)e^{\theta^T X} \Rightarrow P = \frac{1}{1+e^{-\theta^T X}} \end{aligned} \quad (4)$$

4) METHOD-II K-NEAREST NEIGHBORS (KNN)

K nearest neighbors is a straightforward method that maintains all existing examples and categorizes new ones using a similarity metric (e.g., distance functions). KNN has been utilized as a non-parametric approach in statistical estimates and pattern recognition since the early 1970s. KNN (K-Nearest Neighbors) is a classifier technique that is based on ‘‘how similar’’ is a data (a vector) from other [36]. It is one of several (supervised learning) algorithms used in data mining and machine learning.

The KNN’s steps are as follows:

- Get info that isn’t classified;
- Calculate the distance (Euclidian, Manhattan, Murkowski, or Weighted) between the new data and all previously categorized data.
- Calculate the distance (Euclidian, Manhattan, Murkowski, or Weighted) between the new data and all previously categorized data.
 - Examine the list of classes with the lowest distance between them and count how many of each class appear;
 - Takes the class with the most appearances as the proper class;
 - Sorts the new data into the class you choose in step 7;

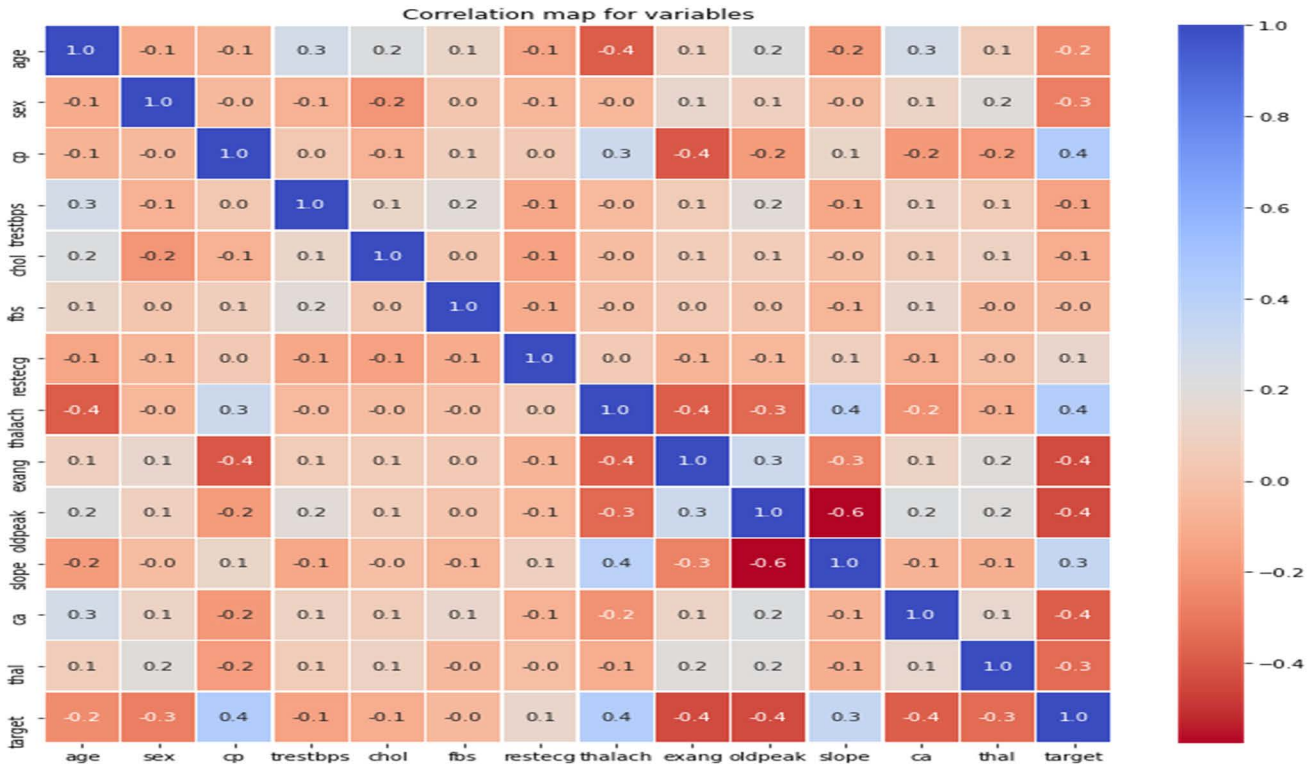


FIGURE 6. Heat map to show correlation.

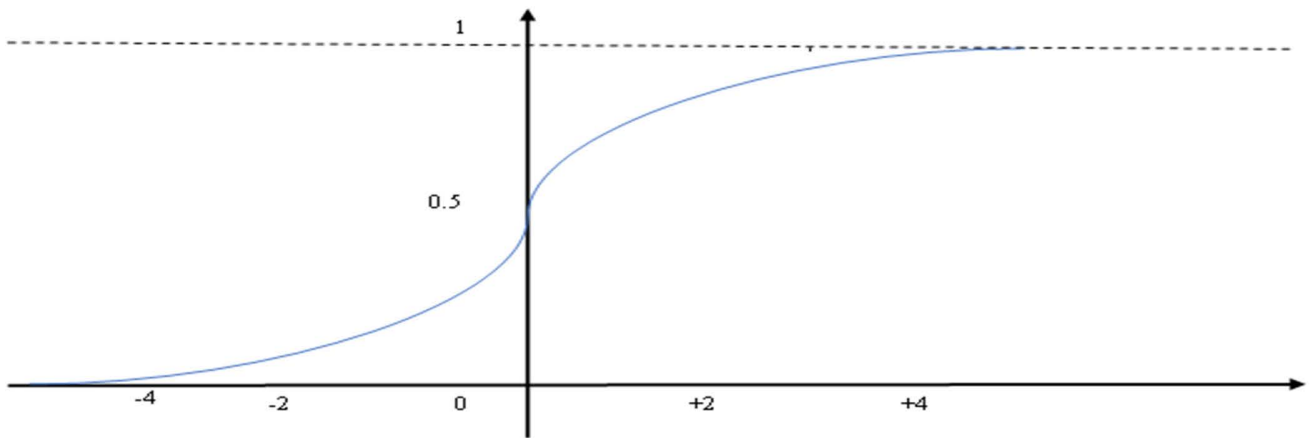


FIGURE 7. Sigmoid function.

Below is a graphic that depicts all of the processes discussed in this post; you have unclassified data (in red) and all of your other data (in yellow and purple), each with your class (A or B). So you calculate the distances between your new data and all the others to see which ones have the smallest distance, and then you take 3 (or 6) of the closest data and see which class appears most frequently. For example, in the image below, the closest data to the new data are those that are inside the first circle (inside circle), and there are 3 other data inside this circle (already classified with yellow), We'll see

which class is the most prevalent; see, it's purple, since there are two purple balls and only one yellow, therefore this new data, which was previously unclassified, will now be classed as purple.

- If we chose $K = 3$, we will have two Class B observations and one Class A observation. As a result, the red star is assigned to Class B.
- If $K = 6$, we have two observations in Class B and four in Class A. As a result, the red star receives a Class A classification.

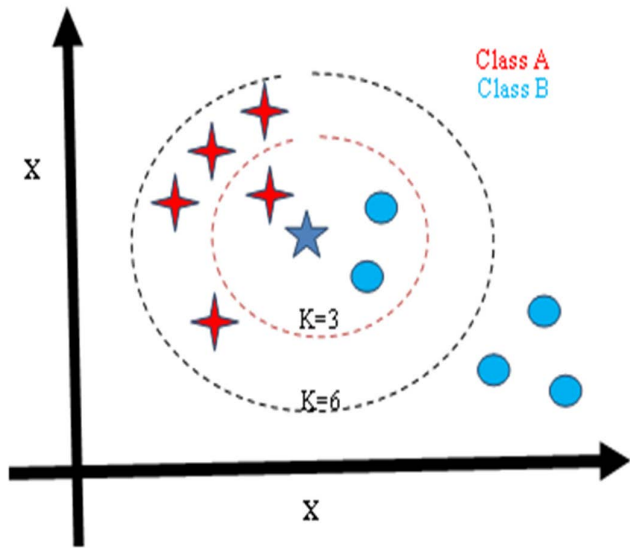


FIGURE 8. k-Nearest neighbors.

Distance calculation: The distance between two points (your new sample and all of the data in your dataset) may be calculated in a variety of ways; as previously said, there are numerous ways to obtain this value; in this article, we will utilize the Euclidean distance [37]. The formula for the Euclidean distance is (5), as shown at the bottom of the page. Given two feature vectors with numeric values $P_i = (p_1, p_2, p_3, \dots, p_n)$ and $Q_i = (q_1, q_2, q_3, \dots, q_n)$.

Training examples (P_i, Q_i) , $p_1, p_2, p_3, \dots, p_n$ Attribute -value represent of example and $q_1, q_2, q_3, \dots, q_n$ real valued target.

5) METHOD-III SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) is a potential machine learning approach based on statistical learning theory. The support vector machine (SVM) is a machine that can classify both linear and nonlinear data. It separates the data into two categories by generating a linear optimum separating Hyperplane inside a higher dimension with the use of support vectors and margins (or classes). The original training data is transferred onto a higher dimension using an appropriate nonlinear mapping. Within this, a hyperplane may always be used to split data from two classes [38]. If f is a classification function for Support vector machines, then

$$f : P \rightarrow Q, \text{ Where } P \text{ is the Domain (here i.e., data set),}$$

$$P = [X, Y], X = [x_i / 1 \leq i \leq n], Y = [y_i / 1 \leq i \leq n] \quad (6)$$

The collection of n training tuples with associated class labels y_i is referred to as x_i , Each y_i can take one of two values: $+1$ or -1 , which correspond to two different classes as: $y_i \in \{+1, -1\}$

The term "output set" can be used to describe a group of results as:

$$Q_i = \{u / 1 \leq i \leq n\} \quad (7)$$

The main principle of SVM is to discover the hyper plane with the greatest margin to distinguish a collection of positive examples from a set of negative examples, as shown in Fig. 9. A linear classification of the type is computed using a support vector machine is as: $f(x) = b + wx$.

The best hyperplane with the most margins, a linear classification of the type is computed using the support vector machine. $f(x) = b + w \cdot x$, where w is a weight vector, x denotes a training sample, and b denotes bias.

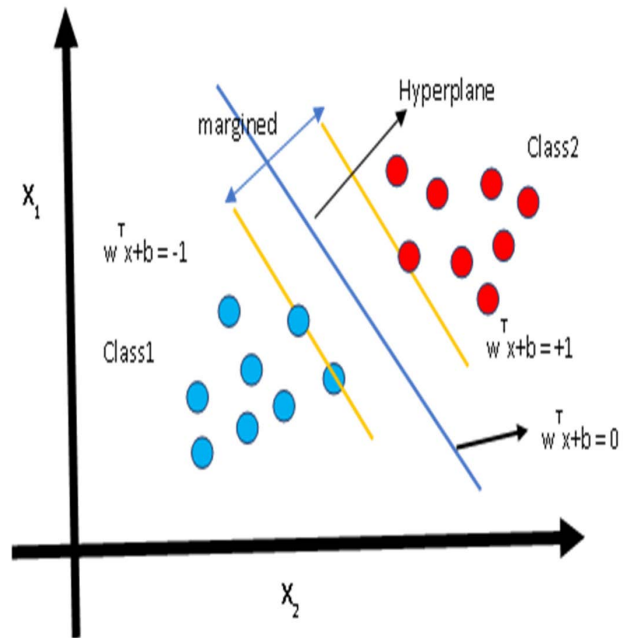


FIGURE 9. Support vector machine.

The hyperplane that separates the two hyperplanes may be expressed as $f(x) = 0$, As a result, any point from one class is located above the dividing hyperplane satisfies, $f(x) > 0$ Similarly, any point from a different class is below the separating hyperplane. satisfies, $f(x) < 0$. The above equations were used to form the linearly separable set, which satisfied the inequality, $y_i (f(x)) \geq 1, \forall i$, Here the margin m is, $m = \frac{1}{\|w\|_2}$

$$d(P_i, Q_i) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + (q_3 - p_3)^2 + \dots + (q_n - p_n)^2}$$

$$d(P_i, Q_i) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (5)$$

TABLE 2. LR model training dataset.

Confusion matrix for using LR for Kaggle's Heart Disease Cleveland, Hungary, Switzerland & Long Beach V Dataset			
Actual Class	Predicted Class		Actual Value
	Absent	Present	
Absent	295	45	340
Present	30	347	377
Total predicted	325	392	

TABLE 3. LR model testing dataset.

Confusion matrix for using LR for Kaggle's Heart Disease Cleveland, Hungary, Switzerland & Long Beach V Dataset			
Actual Class	Predicted Class		Actual Value
	Absent	Present	
Absent	125	34	159
Present	22	127	149
Total predicted	147	161	

TABLE 4. LR model training dataset.

Confusion matrix for using LR for Heart Disease UCI Kaggle data			
Actual Class	Predicted Class		Actual Value
	Absent	Present	
Absent	248	50	298
Present	39	305	344
Total predicted	277	355	

TABLE 5. LR model testing dataset.

Confusion matrix for using LR for Heart Disease UCI Kaggle data			
Actual Class	Predicted Class		Actual Value
	Absent	Present	
Absent	99	13	112
Present	20	144	164
Total predicted	119	157	

Maximizing margin may be represented as an optimization problem using the above equation:

$$\text{Min}_{w,b} \frac{1}{2} \|w\|^2$$

$$\text{Subject to } y_i (w \cdot x + b) \geq 1, \quad \forall i \quad (8)$$

TABLE 6. K-NN model training dataset.

Confusion matrix for using k-NN for Kaggle's Heart Disease Cleveland, Hungary, Switzerland & Long Beach V Dataset			
Actual Class	Predicted Class		Actual Value
	Absent	Present	
Absent	310	30	340
Present	29	348	377
Total predicted	339	378	

TABLE 7. K-NN model testing dataset.

Confusion matrix for using k-NN for Kaggle's Heart Disease Cleveland, Hungary, Switzerland & Long Beach V Dataset			
Actual Class	Predicted Class		Actual Value
	Absent	Present	
Absent	123	36	159
Present	20	129	149
Total predicted	143	165	

TABLE 8. K-NN model training dataset.

Confusion matrix for using K-NN for Heart Disease UCI Kaggle data set			
Actual Class	Predicted Class		Actual Value
	Absent	Present	
Absent	251	47	298
Present	37	313	350
Total predicted	298	350	

TABLE 9. K-NN model testing dataset.

Confusion matrix for using K-NN for Heart Disease UCI Kaggle data V Dataset			
Actual Class	Predicted Class		Actual Value
	Absent	Present	
Absent	97	15	112
Present	16	148	154
Total predicted	113	163	

The dual Lagrange multiplier can be used to tackle this optimization challenge,

$$\text{Min}_{\tilde{\alpha}} \Psi(\tilde{\alpha}) = \text{Min}_{\tilde{\alpha}} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N y_i y_j (\tilde{x}_i \cdot \tilde{x}_j) \alpha_i \alpha_j$$

$$- \sum_{i=1}^N \alpha_i \quad (9)$$

TABLE 10. SVM model training dataset.

Confusion matrix for using SVM for Kaggle's Heart Disease Cleveland, Hungary, Switzerland & Long Beach V Dataset			
Actual Class	Predicted Class		
	Absent	Present	Actual Value
Absent	317	23	340
Present	10	367	377
Total predicted	327	390	

TABLE 11. SVM model testing dataset.

Confusion matrix for using K-NN for Kaggle's Heart Disease Cleveland, Hungary, Switzerland & Long Beach V Dataset			
Actual Class	Predicted Class		
	Absent	Present	Actual Value
Absent	137	22	159
Present	8	141	149
Total predicted	145	163	

TABLE 12. SVM model training dataset.

Confusion matrix for using SVM for Heart Disease UCI Kaggle data V Dataset			
Actual Class	Predicted Class		
	Absent	Present	Actual Value
Absent	256	42	298
Present	22	322	344
Total predicted	278	364	

The support vectors for linearly separable data are a subset of the actual training tuples. A dot product exists between the support vector x_i and the test tuple x_i in the Lagrangian formulation of the aforementioned optimization problem. Each Lagrange multiplier and each training tuple have a one-to-one connection.

Not all data sets can be separated in a linear fashion. There may not be a hyperplane that separates the positive and negative cases. SVMs may also be used to create non-linear classifiers. The Lagrange multiplier is used to compute the output of a non-linear SVM

$$u = \sum_{j=1}^N y_j \alpha_j K(\vec{x}_j, \vec{x}) - b, \quad (10)$$

K is the kernel function. In this case, we employed the Radial Basis Kernel Function (RBF), which is written as follows:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \quad \gamma > 0 \quad (11)$$

TABLE 13. SVM model testing dataset.

Confusion matrix for using SVM for Heart Disease UCI Kaggle data V Dataset			
Actual Class	Predicted Class		
	Absent	Present	Actual Value
Absent	95	17	112
Present	15	149	154
Total predicted	110	166	

TABLE 14. XGBoost model training dataset.

Confusion matrix for using XGBoost for Kaggle's Heart Disease Cleveland, Hungary, Switzerland & Long Beach V Dataset			
Actual Class	Predicted Class		
	Absent	Present	Actual Value
Absent	340	0	340
Present	0	377	377
Total predicted	340	377	

TABLE 15. XGBoost model testing dataset.

Confusion matrix for using XGBoost for Kaggle's Heart Disease Cleveland, Hungary, Switzerland & Long Beach V Dataset			
Actual Class	Predicted Class		
	Absent	Present	Actual Value
Absent	159	0	159
Present	6	143	149
Total predicted	165	143	

The quadratic form is altered by the nonlinearity, but the dual objective function Ψ is still quadratic in,

$$\begin{aligned} \min_{\vec{\alpha}} \Psi(\vec{\alpha}) = & \min_{\vec{\alpha}} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N y_i y_j (\vec{x}_i \cdot \vec{x}_j) \alpha_i \alpha_j \\ & - \sum_{i=1}^N \alpha_i, \quad 0 \leq \alpha_i \leq C, \\ & \forall i \sum_{j=1}^N y_j \alpha_j = 0 \end{aligned} \quad (12)$$

The aforementioned quadratic programming issue is solved using a sequential minimum optimization approach.

6) METHOD-IV XGBoost CLASSIFIER

The gradient boosted trees technique's XGBoost (extreme Gradient Boosting) version is a popular and efficient open-source implementation. Gradient boosting is a supervised learning strategy that attempts to properly forecast a target variable by combining an ensemble of estimates from a variety of simpler and weaker models. The XGBoost approach performs well in machine learning competitions due to its

TABLE 16. XGBoost model training dataset.

Confusion matrix for using XGBoost for Heart Disease UCI Kaggle data V Dataset				
Actual Class	Predicted Class			Actual Value
	Absent	Present		
Absent	298	0		298
Present	0	344		344
Total predicted	298	344		

TABLE 17. XGBoost model testing dataset.

Confusion matrix for using XGBoost for Heart Disease UCI Kaggle data V Dataset				
Actual Class	Predicted Class			Actual Value
	Absent	Present		
Absent	98	14		112
Present	32	132		164
Total predicted	130	146		

TABLE 18. Tuning LR model training dataset.

Confusion matrix for using Tuning LR for Kaggle's Heart Disease Cleveland, Hungary, Switzerland & Long Beach V Dataset				
Actual Class	Predicted Class			Actual Value
	Absent	Present		
Absent	296	44		340
Present	36	341		377
Total predicted	332	385		

TABLE 19. Tuning LR model testing dataset.

Confusion matrix for using Tuning LR for Kaggle's Heart Disease Cleveland, Hungary, Switzerland & Long Beach V Dataset				
Actual Class	Predicted Class			Actual Value
	Absent	Present		
Absent	124	35		159
Present	20	129		149
Total predicted	144	164		

robust handling of a variety of data types, relationships, and distributions, as well as the variety of hyper parameters that can be fine-tuned. Regression, classification (binary and multiclass), and ranking problems may all be solved with XGBoost [39]. The flow chart XGBoost Classifier flow chart is shown in Fig. 10.

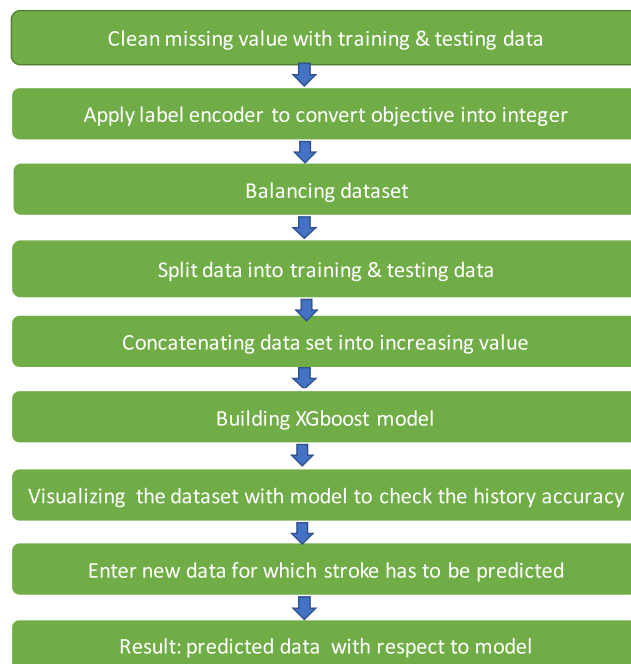


FIGURE 10. XGBoost Classifier flow chart.

TABLE 20. Tuning LR model training dataset.

Confusion matrix for using Tuning LR for Heart Disease UCI Kaggle data V Dataset				
Actual Class	Predicted Class			Actual Value
	Absent	Present		
Absent	248	50		298
Present	39	305		344
Total predicted	287	355		

TABLE 21. Tuning LR model testing dataset.

Confusion matrix for using Tuning for Heart Disease UCI Kaggle data V Dataset				
Actual Class	Predicted Class			Actual Value
	Absent	Present		
Absent	101	11		112
Present	21	143		164
Total predicted	122	154		

7) GRIDSEARCHCV

In virtually all machine learning applications, we train many models on a dataset and then choose the one that performs the best. However, for the sake of progress, we cannot state with certainty that this model is the best for the situation at hand. As a result, our goal is to improve the model in whatever manner we can. One essential feature in these models' performances is that after we select proper values for these hyperparameters, the model's performance can considerably increase. GridSearchCV [40] may also be used to determine

TABLE 22. Tuning K-NN model training dataset.

Confusion matrix for using Tuning K-NN for Kaggle's Heart Disease Cleveland, Hungary, Switzerland & Long Beach V Dataset			
Actual Class	Predicted Class		
	Absence	Presence	Actual Value
Absence	291	49	340
Presence	32	345	377
Total predicted	323	394	

TABLE 23. Tuning K-NN model testing dataset.

Confusion matrix for using Tuning K-NN for Kaggle's Heart Disease Cleveland, Hungary, Switzerland & Long Beach V Dataset			
Actual Class	Predicted Class		
	Absent	Present	Actual Value
Absent	122	37	159
Present	14	135	149
Total predicted	136	172	

TABLE 24. Tuning K-NN model training dataset.

Confusion matrix for using Tuning K-NN for Heart Disease UCI Kaggle data V Dataset			
Actual Class	Predicted Class		
	Absence	Presence	Actual Value
Absence	251	47	298
Presence	35	309	334
Total predicted	286	356	

TABLE 25. Tuning K-NN model testing dataset.

Confusion matrix for using Tuning K-NN for Heart Disease UCI Kaggle data V Dataset			
Actual Class	Predicted Class		
	Absent	Present	Actual Value
Absent	95	17	112
Present	21	143	164
Total predicted	116	160	

optimal values for a model's hyperparameters. Choosing the best hyper parameters have a big impact on the performance model. There are a number of optimization techniques to choose from, each with its own set of advantages and disadvantages. Experiments were done on numerous optimization

TABLE 26. Tuning SVM model training dataset.

Confusion matrix for using Tuning SVM for Kaggle's Heart Disease Cleveland, Hungary, Switzerland & Long Beach V Dataset			
Actual Class	Predicted Class		
	Absent	Present	Actual Value
Absent	340	0	340
Present	0	377	377
Total predicted	340	377	

TABLE 27. Tuning SVM model testing dataset.

Confusion matrix for using Tuning SVM for Kaggle's Heart Disease Cleveland, Hungary, Switzerland & Long Beach V Dataset			
Actual Class	Predicted Class		
	Absent	Present	Actual Value
Absent	159	0	159
Present	6	143	149
Total predicted	165	143	

methodologies to discover the ideal hyper parameter combination, which was then used to Logistic Regression, k-Nearest Neighbors, Support Vector Machine, and XG Boost. The machine learning model has two types of parameters. There are two sorts of parameters: hyperparameters and model parameters. The User must set random hyper parameters before training the Machine Learning Model. The learning process is regulated by the parameters. Different learning rates or weights are used to govern the learning process and find hidden patterns in the data for the same type of machine learning model. To decrease inaccuracy and improve model accuracy, these hyper parameters are fine-tuned. Using a trial-and-error method, these parameters are modified until the ideal hyper parameters are found. Hyper settings that achieve a balance between overfitting and underfitting are optimal. The ability to properly explore the search space is aided by selecting great hyper parameters. To improve the performance of the Logistic Regression, k-Nearest Neighbours, Support Vector Machine, and XG Boost models, the Hyper parameters can be modified. Model parameters are learnt during the training phase. Grid search techniques are commonly used in hyper parameter optimization. Grid Search is a well-known method for determining all hyper parameter combinations. The learning rate and the number of layers are the two most important parameters in Grid Search. A collection of values is initially determined for each hyper parameter. In each cycle, the hyper parameter combination is determined. Finally, the most successful hyper parameter

TABLE 28. Tuning SVM model training dataset.

Confusion matrix for using Tuning SVM for Heart Disease UCI Kaggle data V Dataset				
Actual Class	Predicted Class			Actual Value
	Absent	Present		
Absent	256	42		298
Present	22	322		344
Total predicted	278	364		

TABLE 29. Tuning K-NN model testing dataset.

Confusion matrix for using Tuning SVM for Heart Disease UCI Kaggle data V Dataset				
Actual Class	Predicted Class			Actual Value
	Absent	Present		
Absent	95	17		112
Present	15	149		164
Total predicted	110	166		

TABLE 30. Tuning XGBoost model training dataset.

Confusion matrix for using Tuning XGBoost for Kaggle's Heart Disease Cleveland, Hungary, Switzerland & Long Beach V Dataset				
Actual Class	Predicted Class			Actual Value
	Absent	Present		
Absent	340	0		340
Present	0	377		377
Total predicted	340	377		

TABLE 31. Tuning XGBoost model testing dataset.

Confusion matrix for using Tuning XGBoost for Kaggle's Heart Disease Cleveland, Hungary, Switzerland & Long Beach V Dataset				
Actual Class	Predicted Class			Actual Value
	Absent	Present		
Absent	159	0		159
Present	3	146		149
Total predicted	162	146		

combination is selected and implemented in the learning process.

8) PERFORMANCE MATRIX

The performance of the two models was assessed using a confusion matrix. The count of anticipated values that accurately recognised the presence of diseases is known as True

TABLE 32. Tuning XGBoost model training dataset.

Confusion matrix for using Tuning XGBoost for Heart Disease UCI Kaggle data V Dataset				
Actual Class	Predicted Class			Actual Value
	Absent	Present		
Absent	298	0		298
Present	0	344		344
Total predicted	398	344		

TABLE 33. Tuning XGBoost model testing dataset.

Confusion matrix for using Tuning for Heart Disease UCI Kaggle data V Dataset				
Actual Class	Predicted Class			Actual Value
	Absent	Present		
Absent	99	13		112
Present	27	137		164
Total predicted	126	150		

TABLE 34. Model training result for Kaggle's heart disease Cleveland, Hungary, Switzerland & long beach V dataset.

Model	Training Accuracy (%)	precision (%)	Recall (%)	F1-score (%)
LR	89.54	90	90	90
K-NN	91.77	92	92	92
SVM	95.40	95	95	95
XGB	100	100	100	100

TABLE 35. Model testing results for Kaggle's heart disease Cleveland, Hungary, Switzerland & long beach V dataset.

Model	Testing accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LR	81.82	82	82	82
K-NN	81.82	82	82	82
SVM	90.26	90	90	90
XGB	98.05	98	98	98

Positive (tp). The count of expected values properly identifying the absence of diseases is known as a true negative. The number of anticipated values that are wrongly labelled as

TABLE 36. Tuned model training results for Kaggle’s heart disease Cleveland, Hungary, Switzerland & long beach V dataset.

Model	Training accuracy (%)	precision (%)	recall (%)	F1-score (%)
LR	88.84	89	89	89
K-NN	88.70	89	89	89
SVM	100	100	100	100
XGB	100	100	100	100

TABLE 37. Tuned model testing results for Kaggle’s heart disease Cleveland, Hungary, Switzerland & long beach V dataset.

Model	Testing accuracy (%)	precision (%)	recall (%)	F1-score (%)
LR	82.14	82	82	82
K-NN	83.44	83	83	83
SVM	98.05	98	98	98
XGB	99.03	99	99	99

TABLE 38. Model training result for heart disease UCI Kaggle data V dataset.

Model	Training Accuracy (%)	precision (%)	recall (%)	F1-score (%)
LR	86.14	86	86	86
K-NN	87.23	88	88	88
SVM	90.03	90	90	90
XGB	100	100	100	100

TABLE 39. Model testing results for heart disease UCI Kaggle data V dataset.

Model	Testing accuracy (%)	precision (%)	recall (%)	F1-score (%)
LR	88.04	88	88	88
K-NN	88.77	89	89	89
SVM	88.41	88	88	88
XGB	83.33	83	83	83

positive is known as false positive (fp) (actually when it was negative). The count of anticipated values mistakenly labelled as negative is known as False Negative (fn) (actually when

TABLE 40. Tuned model training results for heart disease UCI Kaggle data V dataset.

Model	Training accuracy (%)	precision (%)	recall (%)	F1-score (%)
LR	86.14	86	86	86
K-NN	87.23	87	87	87
SVM	90.03	90	90	90
XGB	100	100	100	100

TABLE 41. Tuned model testing results for heart disease UCI Kaggle data V dataset.

Model	Testing accuracy (%)	precision (%)	recall (%)	F1-score (%)
LR	88.14	88	88	88
K-NN	86.14	86	86	86
SVM	88.41	88	88	88
XGB	85.51	86	86	86

TABLE 42. Kaggle’s heart disease Cleveland, Hungary, Switzerland & long beach V dataset.

Model	Training accuracy (%)	Testing accuracy (%)
Tuned Logistic Regression	88.84	82.14
Tuned K-nearest neighbours	88.70	83.44
Tuned Support Vector Machine	100.00	98.05
Tuned XGBoost Classifier	100.00	99.03

it was positive). Once the model has been trained, five-fold cross validation is used to predict and verify the risk of heart disease. The investigation employed the performance metrics Accuracy, Recall, Precision, and f-measure score values [22], [41]–[44].

$$Accuracy = \frac{tp + tn}{tp + tn + fn + fp} \tag{13}$$

$$Recall = \frac{tp}{tp + fn} \tag{14}$$

$$Precision = \frac{tp}{tp + fp} \tag{15}$$

$$f - measure = \frac{2 * Recall * Precision}{Recall + Precision} \tag{16}$$

IV. EXPERIMENT OUTCOMES

A. SETUP FOR THE EXPERIMENT

The experiments were carried out. in Anaconda3 2021.05 (Python 3.8.8 64-bit) on Lenovo Intel(R) Core (TM)

TABLE 43. Kaggle’s heart disease Cleveland, Hungary, Switzerland & long beach V dataset.

Model	Training accuracy (%)	Testing accuracy (%)
Logistic Regression	89.54	81.82
K-nearest neighbours	91.77	81.82
Support Vector Machine	95.40	90.26
XGBoost Classifier	100.0	98.05

TABLE 44. The best hyperparameter of XG BOOST with optimization technique for Kaggle’s heart disease Cleveland, Hungary, Switzerland & long beach V dataset.

model	parameter	GridsearchCV	Testing Accuracy (%)
XGB	Learning rate	0.95859476	99.03
	max-depth	8	
	n-estimators	402	

TABLE 45. Heart disease UCI Kaggle data V dataset.

Model	Training accuracy (%)	Testing accuracy (%)
Tuned Logistic Regression	86.14	88.41
Tuned K-nearest neighbours	87.23	86.23
Tuned Support Vector Machine	90.03	88.41
Tuned XGBoost Classifier	100.00	85.51

i3-10110U CPU @ 2.10GHz 2.59 GHz, Intel (R), RAM 8 GB with Windows 10.

B. RESULTS AND ANALYSIS

This study looked at the effect of hyper parameter on the prediction performance of four different machine learning models. The ML models Logistic Regression, k-Nearest Neighbours, Support Vector Machine, and XG Boost were improved using the hyper parameter optimization approach GridSearchCV. In the trials, hyper parameter techniques were used to compare the prediction performance of the four algorithms LR, K-NN, SVM, and XGboost. Each approach is analysed using different hyper parameters. The findings of Randomized Search, Grid Search, and other hyper parameter optimization procedures were compared to those of established approaches. Data was taken for training and testing in proportions of 70% and 30%, respectively. We employed five-fold Cross Validation in our research. Previous research has shown that 5-fold cross validation yields a generalized model while avoiding overfitting.

The presence of heart diseases is represented as one in our study, whereas the absence of heart disease is represented as

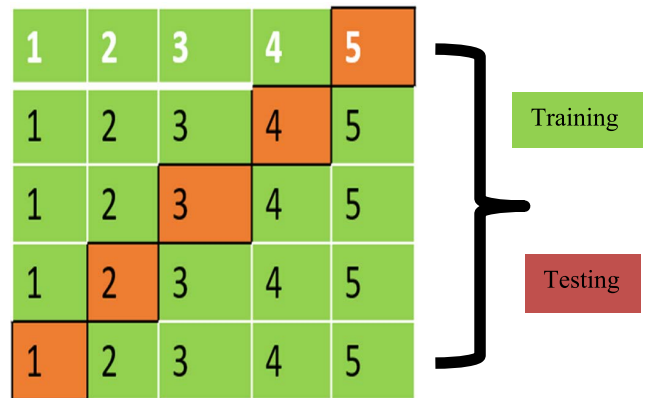


FIGURE 11. Graphical representation of five-fold cross-validation.

TABLE 46. Heart disease UCI Kaggle dataset.

Model	Training accuracy (%)	Testing accuracy (%)
Logistic Regression	86.14	88.04
K-nearest neighbours	87.85	88.77
Support Vector Machine	90.03	88.41
XGBoost Classifier	100.0	83.33

TABLE 47. The best hyperparameter of XG BOOST with optimization technique for heart disease UCI Kaggle dataset.

model	parameter	GridsearchCV	Testing Accuracy (%)
LR	Learning rate(C)	0.0885866	88.41
	solver	liblinear	
SVM	Learning rate(C)	1	88.41
	Gamma	0.1	
	kernel	rbf	

zero. We used the Scikit Python package to implement the three hyper parameter optimization strategies. Scikit-learn has built-in support for hyperparameter tuning approaches.

C. EXPERIMENTAL RESULT WITH FOUR DIFFERENT MACHINE LEARNING MODEL TRAINING & TESTING RESULTS (CLEVELAND, HUNGARY, SWITZERLAND & LONG BEACH V) AND UCI KAGGLE DATA

The training and testing confusion matrix is used to assess the Logistic Regression, k-Nearest Neighbours, Support Vector Machine, and XG Boost algorithms. A 2 × 2 matrix represents the classifier’s correct and incorrect predictions.

Table [2-17] shows the confusion matrix.

The training and testing confusion matrix is used to assess Tuning LR, Tuning KNN, Tuning SVM, and Tuning GBC.

TABLE 48. Comparison with other studies using CH dataset.

Author	Method	Accuracy	Precision	Recall	F-1 Score
Hung et al. [45]	Linear SVM	89.9%	NA	87.0%	NA
	Infinite Latent Feature Selection (ILFS)	90.7%	NA	91.0%	NA
A.K.Garate-Escamila et.al. [46]	Gradient-boosted Tree (GBT)	99.4%	100%	98.7%	NA
	ChiSqSelector + PCA and Logistic Regression	99.4%	97.6%	100%	NA
Proposed Model	Tuned Logistic Regression	88.41%	89%	89%	89%
	Tuned K-nearest Neighbours	83.44%	89%	89%	89%
	Tuned Support Vector Machine	98.05%	100%	100%	100%
	Tuned XGBoost Classifier	100%	100%	100%	100%

TABLE 49. The best hyper parameter of XG BOOST with optimization technique for Kaggle’s heart disease Cleveland, Hungary, Switzerland & long beach V dataset.

model	parameter	GridsearchCV	Testing Accuracy (%)
XGB	Learning rate	0.95859476	99.03
	max-depth	8	
	n-estimators	402	

TABLE 50. The best hyperparameter of XG BOOST with optimization technique for heart disease UCI Kaggle dataset.

model	parameter	GridsearchCV	Testing Accuracy (%)
LR	Learning rate(C)	0.0885866	88.41
	solver	liblinear	
SVM	Learning rate(C)	1	88.41
	Gamma	0.1	
	kernel	rbf	

A 2 × 2 matrix represents the classifier’s correct and erroneous predictions. **Table [18-33]** shows the confusion matrix using a five-fold cross-validation method.

Table [34-41] shows the experimental results of four alternative models for training and testing accuracy, precision, recall, and F1-score with and without the GridSearchCV optimization approach.

V. RESULT OF CLASSIFIERS (FIVE-FOLD CROSS VALIDATION) WITH ALL FEATURES (n = 13) AND WITH GRIDSEARCHCV FOR KAGGLE’S HEART DISEASE CLEVELAND, HUNGARY, SWITZERLAND & LONG BEACH V DATASETS

Through GridSearchCV with five-fold cross validation technology, all characteristics of the data set are concentrated

on four machine learning classifiers in this examination. 70% of the fivefold CV was utilised to train the classifier, whereas only 30% was assessed. Finally, the five-fold technique’s normal measurement result is obtained. A number of boundary assessments have also passed the classifier. The Extreme Gradient Boosting Classifier using GridSearchCV performs well with an accuracy of 99.03% in Tables 42&43 and Figure 10 (Kaggle’s Heart Disease Cleveland, Hungary, Switzerland, and Long Beach V Dataset). The “support vector machine with GridSearchCV” accuracy of 98.05% is the following number. The classifiers Tuned Logistic Regression, Tuned K-nearest Neighbours, and Tuned XGBoost Classifier have escalating accuracy of 82.14%, 83.44 %, 98.05 %, and 99.03 %, respectively, if we discriminate between classifiers that use GridSearchCV and those that don’t.

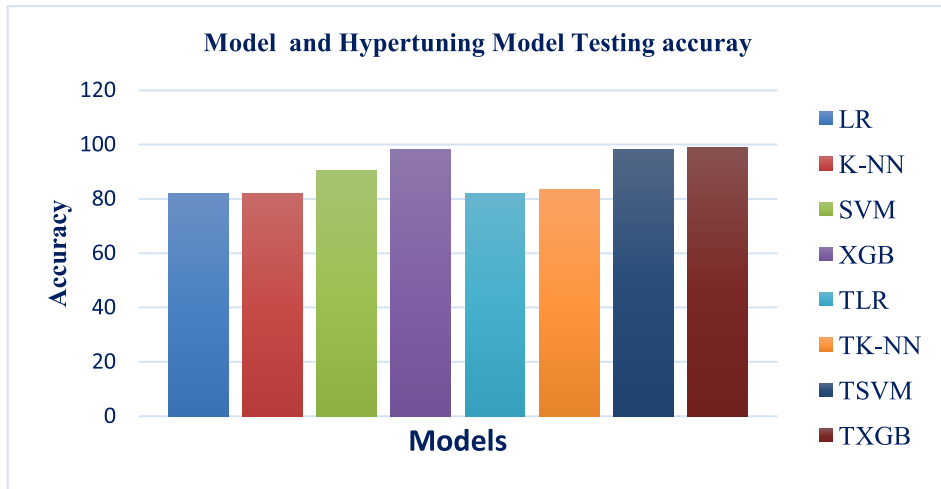


FIGURE 12. Model and hypertuning model testing accuracy.

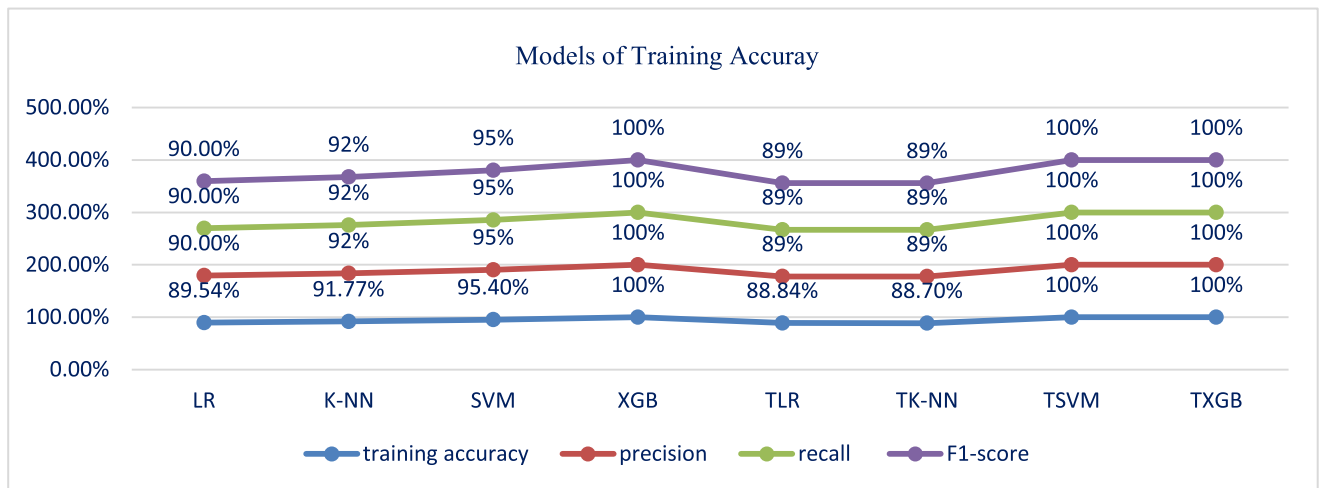


FIGURE 13. Model of training accuracy.

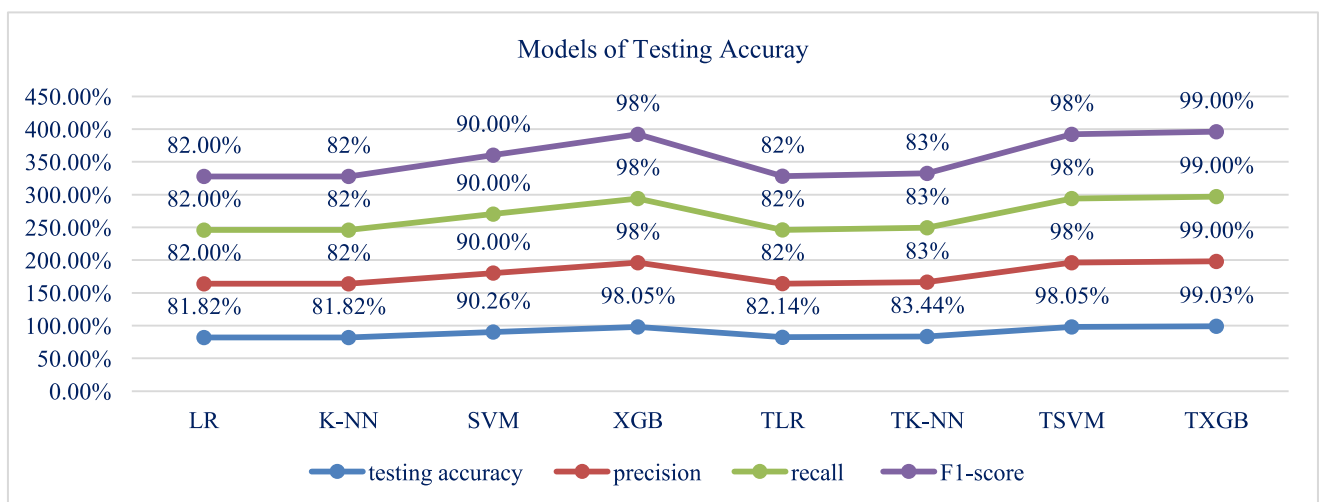


FIGURE 14. Model of testing accuracy.

The classifiers without GridSearchCV, Logistic Regression, K-nearest neighbours, Support Vector Machine, and

XGBoost Classifier, on the other hand, have escalating accuracy of 81.82%, 90.26%, and 98.05 %, respectively. Except

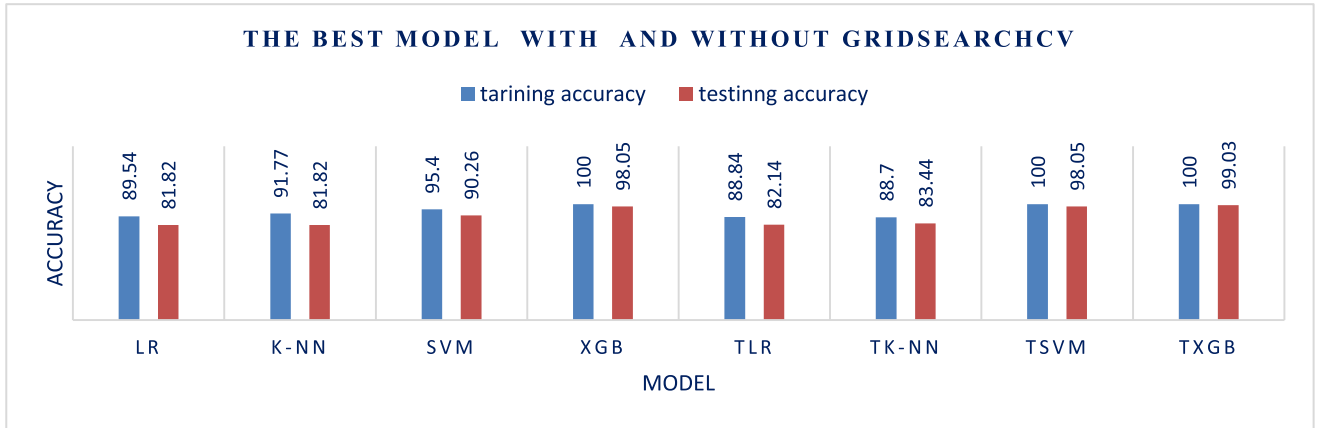


FIGURE 15. The best model with and without GridSearchCV.

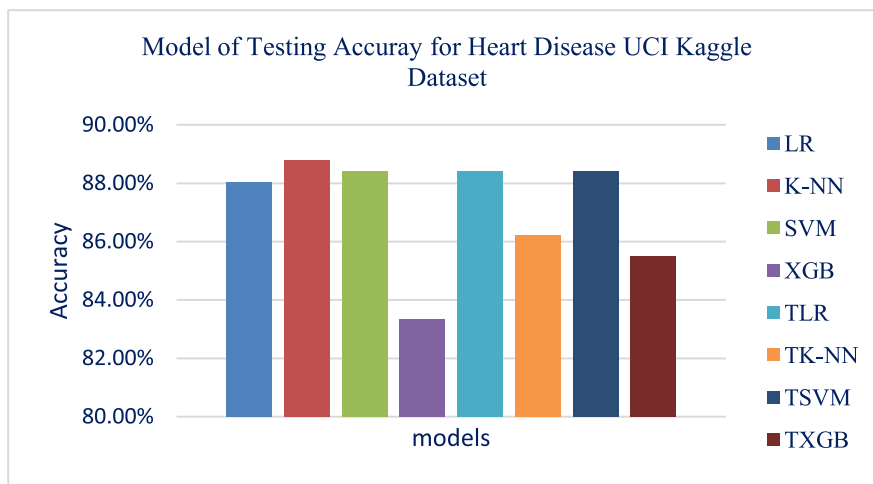


FIGURE 16. Model of testing accuracy with and without GridSearchCV.

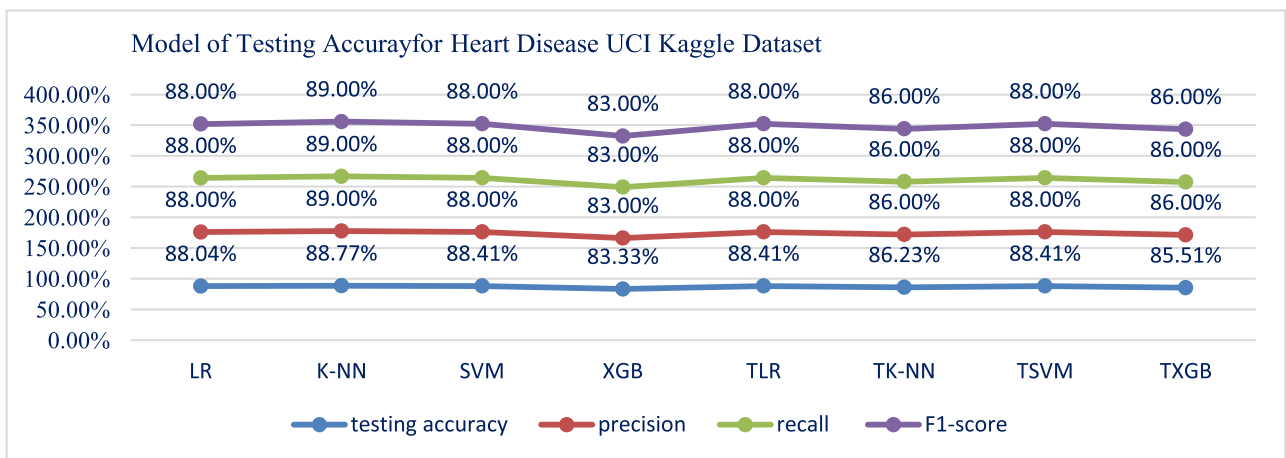


FIGURE 17. Model of testing accuracy with and without GridSearchCV precision, recall, F-1score.

for The Gradient Boosting Classifier with GridSearchCV, which is the best hyperparameter for measuring accuracy,

as shown in Table-44, only The Gradient Boosting Classifier and Support Vector Machine have good invisibility for all

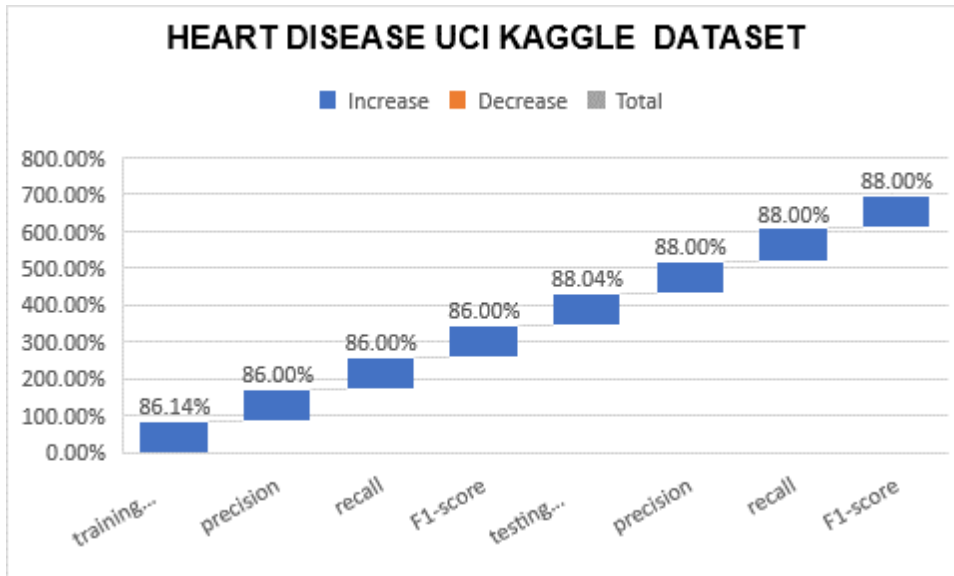


FIGURE 18. Model of training & testing accuracy with and without GridSearchCV precision, recall, F-1score.

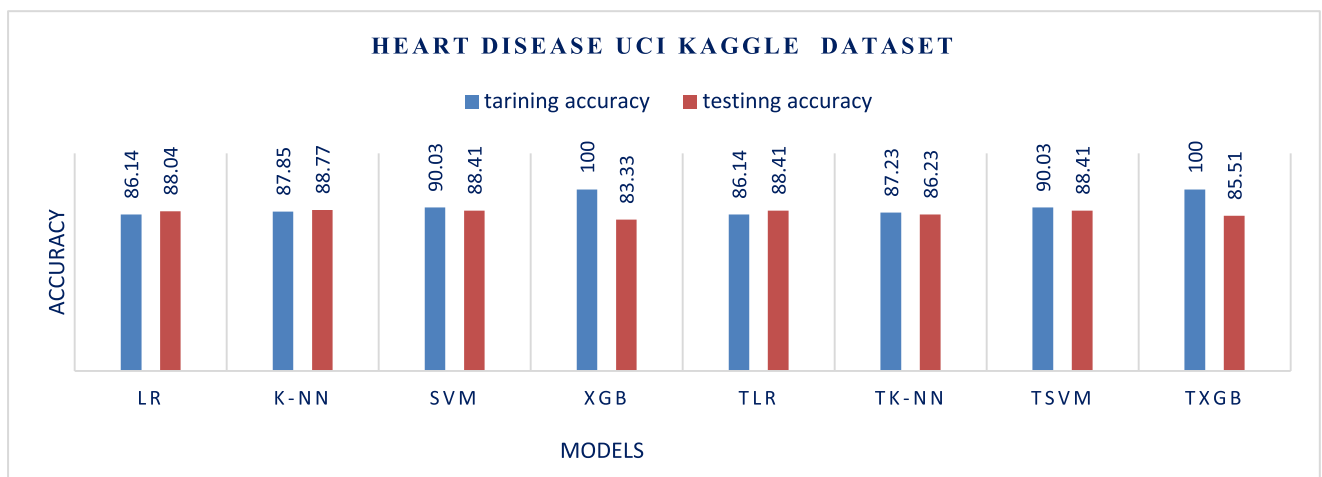


FIGURE 19. Model of training & testing accuracy with and without GridSearchCV.

features, at the very least. In the same sequence as the training and test data sets, Fig. 13 demonstrates the performance of several classifiers.

A. RESULTS OF VALIDATION

B. RESULT OF CLASSIFIERS (FIVE-FOLD CROSS VALIDATION) WITH ALL FEATURES (n = 11) AND WITH GRIDSEARCHCV FOR HEART DISEASE UCI KAGGLE DATA V DATASET

Through GridSearchCV with fivefold cross-validation technology, all characteristics of the data set are concentrated on four machine learning classifiers in this examination. Only 30% of the fivefold CV was examined, whereas 70% was utilized to train the classifier. Finally, the fivefold technique’s normal measurement result is obtained. A number of boundary assessments have also passed the classifier.

In Table-45 & 46 (Heart Disease UCI Kaggle data V Dataset) Using an accuracy of 88.41%, the Logistic Regression with GridSearchCV performs well. The “support vector machine with GridSearchCV” accuracy of 88.41% is the following number. When we compare classifiers with and without GridSearchCV, the accuracy of Tuned Logistic Regression, Tuned K-nearest Neighbors, Tuned Support Vector Machine, and Tuned XGBoost Classifier is 88.41%, 86.23%, 88.41%, and 85.51%, respectively. GridSearchCV, Logistic Regression, K-nearest Neighbors, Support Vector Machine, and XGBoost Classifier, on the other hand, have accuracy of 88.04%, 88.7%, 88.41%, and 83.33%, respectively. Except for The Gradient Boosting Classifier with GridSearchCV is the best Boosting Classifier hyperparameter for testing accuracy show Table-47, only Gradient Classifier and Support Vector Machine have good invisibility for all features, but at least.

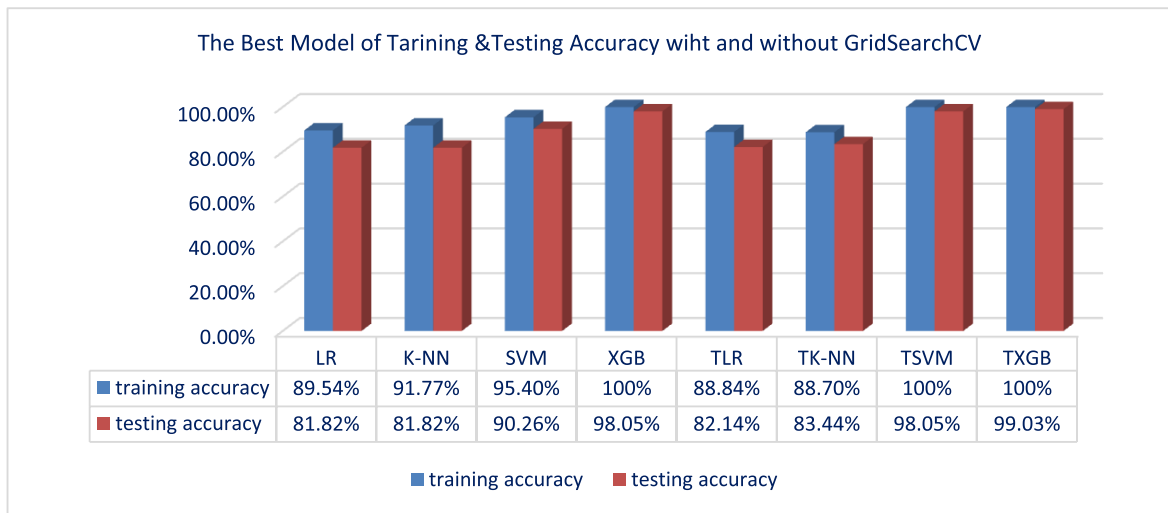


FIGURE 20. The best model of training and testing accuracy with and without GridSearchCV.

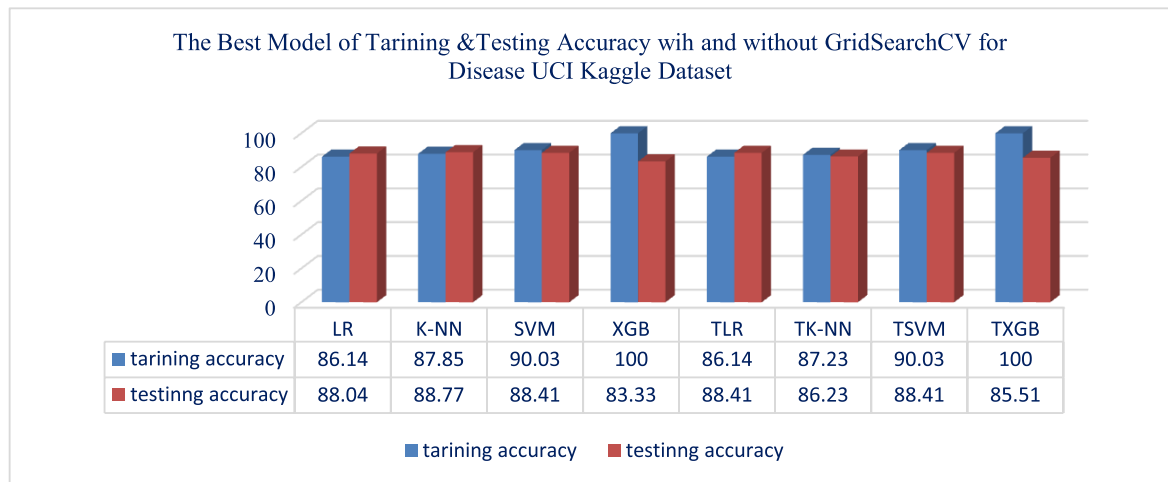


FIGURE 21. The best model of training and testing accuracy with and without GridSearchCV precision, recall, F-1score.

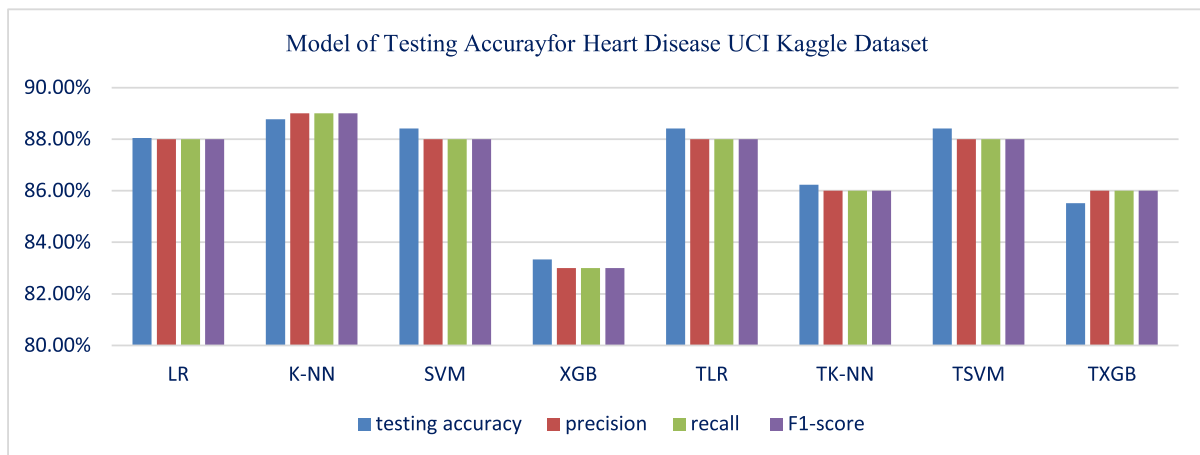


FIGURE 22. The best model of training and testing accuracy with and without GridSearchCV.

Figure 16-19 shows the performance of different classifiers on the training and test data sets in the same order.

Table-48 compares combined heart disease (CHD) dataset's performance to that of the previously published

works. Infinite Latent Feature Selection (ILFS), Sort features according to pairwise correlations (CFS), Feature Selection and Kernel Learning for Local Learning-Based Clustering (LLCFS), and PCA were used by Hung et al. [45]. ILFS

was the classifier that employed the fewest features and performed the best computation with 90.65% accuracy. CFS, LLCFS, and PCA calculated an accuracy of 89.93 %, but they needed at least 55 features to get there. They used the Cleveland, Hungarian, and Swiss databases. With a lesser number of features, we have obtained the improved results as: accuracies, Precision, Recall, F-1 Score for the combined heart disease (CHD) datasets.

VI. CONCLUSION

Predicting heart disease is regarded as one of the most difficult challenges in medicine. Figuring out what's causing this takes a lot of time and work, especially for doctors and other medical specialists. To predict cardiac diseases, a variety of algorithms, including LR, KNN, SVM, GBC, and the GridSearchCV, are used in this paper. For verification, a five-fold cross-validation approach is employed. These four approaches are used in the comparative study. The Datasets for Cleveland, Hungary, Switzerland, and Long Beach V and the Dataset Heart Disease UCI Kaggle are used to assess the models' performance. For both the Hungary, Switzerland & Long Beach V and Heart Disease UCI Kaggle Datasets, the Extreme Gradient Boosting Classifier with GridSearchCV produced the highest and roughly identical accuracy results of 100% and 99.03%, respectively. The Extreme Gradient Boosting Classifier with GridSearchCV is the best hyperparameter for testing accuracy. The results are compared to those of previous cardiac prediction studies. In this study, four algorithms are tuned, trained, and tested with four different models, including logistic regression, K-Nearest Neighbors, support vector machine, and Extreme Gradient Boosting, both with and without hyperparameter tuning techniques. Their accuracies are compared with the existing methods. By modifying the parameters of XGB with Grid Search testing accuracy obtained is 99.03%. Extreme Gradient Boosting Classifier with GridSearchCV has the greatest testing accuracy of these four approaches, at 99.03%. Without GridSearchCV, the Extreme Gradient Boosting Classifier has a testing accuracy of 98.05%. It is evident from table-45 & 47, figures 20 to 22, the Extreme Gradient Boosting Classifier with GridSearchCV is the best hyperparameter for testing accuracy.

In the future, we'd like to improve the model so that it may be used with several feature selection algorithms. Another possibility is to utilize a GridSearchCV with a Gradient Boosting Classifier. The major goal of this study is to improve on previous work by developing a new and unique model-creation method and to make the model relevant and easy to use in a real-world situation.

REFERENCES

- [1] P. Drotár and Z. Smékal, "Comparative study of machine learning techniques for supervised classification of biomedical data," *Acta Electrotechnica Inf.*, vol. 14, no. 3, pp. 5–10, Sep. 2014, doi: [10.15546/aeci-2014-0021](https://doi.org/10.15546/aeci-2014-0021).
- [2] A. Levin, "The clinical epidemiology of cardiovascular diseases in chronic kidney disease: Clinical epidemiology of cardiovascular disease in chronic kidney disease prior to dialysis," in *Seminars in Dialysis*, vol. 16, no. 2. Oxford, U.K.: Blackwell Science, Mar. 2003, pp. 101–105.
- [3] K. S. Reddy, "Cardiovascular diseases in the developing countries: Dimensions, determinants, dynamics and directions for public health action," *Public Health Nutrition*, vol. 5, no. 1, pp. 231–237, Feb. 2002.
- [4] A. Kishore, A. Kumar, K. Singh, M. Punia, and Y. Hambir, "Heart attack prediction using deep learning," *Int. Res. J. Eng. Technol.*, vol. 5, no. 4, p. 2395, 2018.
- [5] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS Med.*, vol. 3, no. 11, p. e442, Nov. 2006.
- [6] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Heart disease prediction system using associative classification and genetic algorithm," in *Proc. Int. Conf. Emerg. Trends Elect., Electron. Commun. Technol. (ICECIT)*, 2012, pp. 40–46.
- [7] T. N. Sugathan, C. R. Soman, and K. Sankaranarayanan, "Behavioural risk factors for non-communicable diseases among adults in Kerala, India," *Indian J. Med. Res.*, vol. 127, no. 6, pp. 1–9, 2008.
- [8] A. Ahmed and S. A. Hannan, "Data mining techniques to find out heart diseases: An overview," *Int. J. Innov. Technol. Exploring Eng.*, vol. 1, no. 4, pp. 18–23, 2012.
- [9] M. Ribeiro, K. Grolinger, and M. A. M. Capretz, "MLaaS: Machine learning as a service," in *Proc. IEEE 14th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2015, pp. 896–902.
- [10] I. Castelli and E. Trentin, "Combination of supervised and unsupervised learning for training the activation functions of neural networks," *Pattern Recognit. Lett.*, vol. 37, pp. 178–191, Feb. 2014.
- [11] Z. Sani, R. Alizadehsani, J. Habibi, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, F. Khozimeh, and F. Alizadeh-Sani, "Diagnosing coronary artery disease via data mining algorithms by considering laboratory and echocardiography features," *Res. Cardiovascular Med.*, vol. 2, no. 3, p. 133, 2013.
- [12] D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare," *Int. J. Bio-Sci. Bio-Technol.*, vol. 5, no. 5, pp. 241–266, 2013.
- [13] Y. Er, "The classification of white wine and red wine according to their physicochemical qualities," *Int. J. Intell. Syst. Appl. Eng.*, vol. 4, no. 1, pp. 23–26, Dec. 2016.
- [14] S. J. Pasha and E. S. Mohamed, "Novel feature reduction (NFR) model with machine learning and data mining algorithms for effective disease risk prediction," *IEEE Access*, vol. 8, pp. 184087–184108, 2020.
- [15] D. Swain, S. K. Pani, and D. Swain, "A metaphoric investigation on prediction of heart disease using machine learning," in *Proc. Int. Conf. Adv. Comput. Telecommun. (ICACAT)*, Bhopal, India, Dec. 2018, pp. 1–6.
- [16] S. F. Weng, J. Repts, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" *PLoS ONE*, vol. 12, no. 4, Apr. 2017, Art. no. e0174944.
- [17] Y. Khan, U. Qamar, N. Yousaf, and A. Khan, "Machine learning techniques for heart disease datasets: A survey," in *Proc. 11th Int. Conf. Mach. Learn. Comput. (ICMLC)*, Zhuhai, China, 2019, pp. 27–35.
- [18] S. Goel, A. Deep, S. Srivastava, and A. Tripathi, "Comparative analysis of various techniques for heart disease prediction," in *Proc. 4th Int. Conf. Inf. Syst. Comput. Netw. (ISCON)*, Mathura, India, Nov. 2019, pp. 88–94.
- [19] V. Chaurasia and S. Pal, "Early prediction of heart diseases using data mining techniques," *Caribbean J. Sci. Technol.*, vol. 1, pp. 208–217, 2013.
- [20] R. Alizadehsani, M. J. Hosseini, Z. A. Sani, A. Ghandeharioun, and R. Boghrati, "Diagnosis of coronary artery disease using cost-sensitive algorithms," in *Proc. IEEE 12th Int. Conf. Data Mining Workshops*, Dec. 2012, pp. 9–16.
- [21] A. Lakshmanarao, Y. Swathi, and P. S. S. Sundareswar, "Machine learning techniques for heart disease prediction," *Int. J. Sci. Technol. Res.*, vol. 8, no. 11, p. 97, Nov. 2019.
- [22] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.
- [23] E. O. Olaniyi, O. K. Oyedotun, and K. Adnan, "Heart diseases diagnosis using neural networks arbitration," *Int. J. Intell. Syst. Appl.*, vol. 7, no. 12, p. 72, 2015.

- [24] N. R. Ratnasari, A. Susanto, I. Soesanti, and Maesadji, "Thoracic X-ray features extraction using thresholding-based ROI template and PCA-based features selection for lung TB classification purposes," in *Proc. 3rd Int. Conf. Instrum., Commun., Inf. Technol. Biomed. Eng. (ICICI-BME)*, Bandung, Indonesia, Nov. 2013, pp. 65–69.
- [25] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *Amer. J. Cardiol.*, vol. 64, no. 5, pp. 304–310, 1989.
- [26] M. Mohana, A. S. Saidi, S. Alelyani, M. J. Alshayeb, S. Basha, and A. E. Anqi, "Small-scale solar photovoltaic power prediction for residential load in Saudi Arabia using machine learning," *Energies*, vol. 14, no. 20, p. 6759, Oct. 2021, doi: [10.3390/en14206759](https://doi.org/10.3390/en14206759).
- [27] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrés, "Classification models for heart disease prediction using feature selection and PCA," *Informat. Med. Unlocked*, vol. 19, Jan. 2020, Art. no. 100330.
- [28] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–11, Jul. 2021.
- [29] D. Zhang, L. Zou, X. Zhou, and F. He, "Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer," *IEEE Access*, vol. 6, pp. 28936–28944, 2018.
- [30] [Online]. Available: <https://www.kaggle.com/johnsmith88/heart-disease-dataset>
- [31] [Online]. Available: <https://archive.ics.uci.edu/ml/dataset/heart+disease>
- [32] D. W. Hosmer, S. Lemeshow, and E. D. Cook, *Applied Logistic Regression*, 2nd ed. New York, NY, USA: Wiley, 2000.
- [33] D. Böhning, "Multinomial logistic regression algorithm," *Ann. Inst. Stat. Math.*, vol. 44, no. 1, pp. 197–200, Oct. 1992.
- [34] M. Steinbach and P. N. Tan, "KNN: K-nearest neighbors," in *The Top Ten Algorithms in Data Mining*. Boca Raton, FL, USA: CRC Press, 2009, pp. 165–176.
- [35] I. Dokmanic, R. Parhizkar, J. Ranieri, and M. Vetterli, "Euclidean distance matrices: Essential theory, algorithms, and applications," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 12–30, Nov. 2015.
- [36] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, Oct. 1995.
- [37] E. Nasarian, M. Abdar, M. A. Fahami, R. Alizadehsani, S. Hussain, M. E. Basiri, M. Zomorodi-Moghadam, X. Zhou, P. Pławiak, U. R. Acharya, R.-S. Tan, and N. Sarrafzadegan, "Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach," *Pattern Recognit. Lett.*, vol. 133, pp. 33–40, May 2020.
- [38] R. Atallah and A. Al-Mousa, "Heart disease detection using machine learning majority voting ensemble method," in *Proc. 2nd Int. Conf. new Trends Comput. Sci. (ICTCS)*, Oct. 2019, pp. 1–6.
- [39] A. Gupta, L. Kumar, R. Jain, and P. Nagrath, "Heart disease prediction using classification (naive bayes)," in *Proc. 1st Int. Conf. Comput., Commun., Cyber-Secur. (ICS)*. Singapore: Springer, 2020, pp. 561–573.
- [40] F. M. J. M. Shamrat, P. Ghosh, M. H. Sadek, M. A. Kazi, and S. Shultana, "Implementation of machine learning algorithms to detect the prognosis rate of kidney disease," in *Proc. IEEE Int. Conf. Innov. Technol.*, Nov. 2020, pp. 1–7.
- [41] F. Miao, Y.-P. Cai, Y.-X. Zhang, X.-M. Fan, and Y. Li, "Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest," *IEEE Access*, vol. 6, pp. 7244–7253, 2018.
- [42] L. Wang, W. Zhou, Q. Chang, J. Chen, and X. Zhou, "Deep ensemble detection of congestive heart failure using short-term RR intervals," *IEEE Access*, vol. 7, pp. 69559–69574, 2019.
- [43] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, "MIFH: A machine intelligence framework for heart disease diagnosis," *IEEE Access*, vol. 8, pp. 14659–14674, 2020.
- [44] P. Ghosh, S. Azam, M. Jonkman, A. Karim, F. M. J. M. Shamrat, E. Ignatious, S. Shultana, A. R. Beeravolu, and F. De Boer, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: [10.1109/ACCESS.2021.3053759](https://doi.org/10.1109/ACCESS.2021.3053759).
- [45] M. H. Le, T. D. Tran, and L. V. Tran, "Automatic heart disease prediction using feature selection and data mining tech," *J. Comput. Sci. Cybern.*, vol. 34, no. 1, pp. 33–48, 2018.
- [46] A. K. Gárate-Escamila, A. H. El Hassani, and E. Andrés, "Classification models for heart disease prediction using feature selection and PCA," *Informat. Med. Unlocked*, vol. 19, Jan. 2020, Art. no. 100330.



GHULAB NABI AHMAD received the bachelor's degree in science and the master's degree from Avadh University, in 2003 and 2005, respectively, and the M.Tech. degree in operation research from the National Institute of Technology (NIT) Durgapur, in 2010. He is currently pursuing the Ph.D. degree with Mangalayatan University, Aligarh, Uttar Pradesh. From August 2010 to January 2018, he joined the Shivdan Singh Institute of Technology and Management, Aligarh, as an Assistant Professor teaching the B.Tech., B.Parma., M.Tech., M.Pharma., and M.B.A. students. His research interests include machine learning techniques, operation research, advance statistics, numerical method, and bio-statistics.



HIRA FATIMA received the Ph.D. degree in pure mathematics from Aligarh Muslim University, Aligarh, India. She is currently working as an Assistant Professor with the Department of Mathematics, Institute of Applied Science, Mangalayatan University, Aligarh. She taught so many courses of mathematics in UG and PG level. Her research interest includes functional analysis & sequence spaces. She has been working on ideal convergence of double sequence. Her 18 research papers has published in national and international journals, including Cogent Mathematics, Filomat, *Journal of Intelligent and Fuzzy Systems*, *Advances in Difference Equations*, *Journal of Mathematical Analysis*, and *Journal of Mathematical & Computer Science*. She has published a book titled *A Study of Triple Sequence Spaces* in Rating Academy. She has also reviewed research papers for the national and international journals, such as *Mathematical Reviews*, *Complexity* journal (Hindawi), and *Journal of Intelligent and Fuzzy Systems*. She is a Life Time Member of Indian Mathematical Society and the Institute of Engineering Research and Publication.



SHAFIULLAH received the bachelor’s degree in statistics and the master’s and Ph.D. degrees in operations research from the Department of Statistics and Operations Research, Aligarh Muslim University (AMU), Aligarh, India, in 2008, 2010, and 2014, respectively. He has about ten years of teaching and research experience, and has been working as an Assistant Professor with the Department of Mathematics, KCTC College Raxaul, Bihar, India. He has authored or coauthored several

research papers in international journals and conference proceedings. His research interests include geometric programming, mathematical programming in sample surveys, integer optimization problems, and optimization in machine learning techniques.



IMDADULLAH (Senior Member, IEEE) received the bachelor’s degree in electrical engineering and the master’s and Ph.D. degrees in power systems and drives from the Department of Electrical Engineering, Zakir Hussain College of Engineering and Technology, Aligarh Muslim University (AMU), Aligarh, India, in 2003, 2006, and 2020, respectively. He has about 15 years of teaching and research experience. Since December 2007, he has been working as an Assistant Professor with the

Electrical Engineering Section, University Polytechnic, AMU. He has published several research articles in international journals/conference proceedings, including articles in IEEE TRANSACTIONS and IEEE journals. He holds a patent on the FASAL System, “A concept of flexible asynchronous AC link.” He is also holding an international patent granted by Federal Republic of Germany. His research interests include renewable energy, power systems, drives, and microgrid integration and control.

...



ABDELAZIZ SALAH SAIDI (Member, IEEE) was born in Tunisia, in 1979. He received the M.S. degree in electrical systems and the Ph.D. degree in electrical engineering from the National Engineering School of Tunis, in 2003 and 2011, respectively. He joined the Higher Institute of Computer Science of Kef as an Assistant Professor, in 2012. He is currently an Associate Professor in electrical engineering with King Khalid University, Abha. His research interests include

dynamical systems theory applied to power systems integrating renewable energy, power systems voltage stability using load flow techniques, and power systems control and operation.