Received: March 21, 2023. Revised: July 18, 2023.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Experimental Analysis of Heart Disease Prediction Using Machine Learning with Emphasis on Hyper Parameter Tuning and Recursive Feature Elimination

Snehal Bankatrao Shinde[1] **Kankipati Lahari**[2] **Keerthika Chowdary Garimella**[2]
Vicharapu Sowmya Sree[2] **Nileshchandra K Pikle**[1] **Girish S Bhavekar**[3]
Pradnya Borkar[4]* **Sagarkumar Badhiye**[4] **Mukesh Raghuwanshi**[4]

[1]*School of Computer Science and Engineering, IIIT Nagpur, India*
[2]*School of Computer Science and Engineering, VIT-AP University, Amaravati, India*
[3]*Department of AI and DS,Chatrapati Shahu Maharaj College of Engineering, Aurangabad, India*
[4]*Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, India*
*Corresponding Authors: pradnyaborkar2@gmail.com

Abstract: Cardiovascular disease is any illness that makes the heart work less well. Researchers are working on making smart systems that can correctly detect heart conditions from electronic health data using machine learning algorithms. This is because heart conditions can be very serious. In this study, data from patients and key clinical factors are used to identify cardiovascular disease using machine learning. The main goal of the suggested model is to improve the accuracy and reliability of predicting cardiac disease by focusing on parameter tuning, ensemble methods, and recursive feature removal approaches. Our methods for making predictions included logistic regression, decision trees, K-nearest neighbour (KNN), support vector machine (SVM), naive bayes (NB) machine learning (ML) approaches, ensemble technique approaches, and artificial neural networks (ANN) with stress on regularisation. Compared to the other ways, it was found that using a KNN model gave the most accurate results for the model. A number of factors, such as accuracy, precision, memory, and F1-score, were used to judge the models. The KNN model is the most accurate, at 97.8%.

Keywords: Cardiovascular disease, Hyper parameter tuning, Machine learning, Artificial neural network, Recursive feature elimination.

1. Introduction

The heart is a vital organ. It controls blood circulation. Cardiac abnormalities can cause bodily pain. Heart disease is any condition that impairs heart function. Today, heart disease kills most people. Tobacco, alcohol, a high-fat diet, and inactivity increase the risk of cardiovascular disease. According to the World health organization (WHO), heart disease kills approximately 10 million people a year. Preventing heart disease requires a healthy lifestyle and early detection. Diagnosis and therapy are the main concerns of modern healthcare. Heart disease is the largest preventable cause of death worldwide. Illness detection determines illness treatment. The proposed approach detects cardiac issues early to avoid negative effects. The application of machine learning in healthcare settings is now receiving a lot of attention due to the fact that it has the potential to enhance both knowledge and, as a result, patient outcomes. Through ML, the accurate detection of a broad variety of illnesses has been made easier tension due to the fact that it has the potential to enhance both knowledge and, as a result, patient outcomes. Through ML, the accurate detection of a broad variety of illnesses has been made easier. Future research supported by powerful and numerous machine learning calculations has the potential to make it possible to accurately forecast the progression of infection and treat patients. The field of health care creates a flood of data concerning medical services on a daily basis. This data may be mined for information regarding illness patterns and the results of therapy. The patient's medical records contain private information that will, at some point in the future, be utilized in order to give the patient the impression that they are playing a more active role in the choices that are made regarding their health. In a similar vein, this region requires growth in the form of the utilization of educational data in the medical field. Artificial intelligence combs through massive amounts of clinical data, learns from it, and anticipates and predicts it, all while providing professional support. This is made possible by cutting-edge computation. The most common applications for machine learning methods include medical research into illnesses, cardiovascular issues, and sensory concerns. It is possible for self-prepared frameworks to follow in the footsteps of both facilitated and unfacilitated learning if early discovery and analysis are made easier through facilitation. There is a clear connection between human motion and machine learning due to the fact that self-prepared frameworks require continual collaboration with information from clinical testing in order to operate at their best. More ways should be looked into to link ML and DL models that have been trained on heart disease to specific multimedia for patients and doctors [42-44]. This will result in an improvement in patient care as well as a reduction in problems. It is necessary to use ML in order to analyze data and discover new discrete patterns [41] [46]. It provides data analysis and makes predictions about the risk of heart disease. The authors tested various machine learning and deep learning methods for the goal of identifying heart disease. To increase the precision of their findings, they engaged in hyperparameter tuning. Neural networks outperformed other models like logistic regression, SVM, and ensemble techniques like random forest with a high accuracy of 89% [1]. Using the accuracy-based weighted average classifier ensemble technique, an ensemble is created by first representing the various groups using the classification and regression trees (CART) method. At 91% accuracy, it does better than KNN, logistic regression (LR), linear discriminant analysis, SVM, decision trees, gradient boosts, and random forests (RF) [2].

The discussion above makes it evident that the accuracy found in individual research projects is currently unsatisfactory. Compared to other algorithms, some of them offer better performance. The objective of this research work is to recognize classifiers that can accurately predict heart disease sufficiently to be of use in clinical settings. In this study, the data set is retrieved from the Kaggle website. The early cardiac disease prediction is investigated using a variety of statistical models, including Naive Bayes, decision trees, LR, SVM, and KNN. The ensemble approaches such as bagging, boosting, stacking, and random forest have been used to predict cardiovascular disease based on patient features and data. In this experiment, KNN helped to achieve more precision.

This study of research is discussed in various sections as follows: Section 2 offers a comprehensive work pertinent to the current research, and section 3 includes prerequisites used to understand the proposed work along with the Methodology. Lastly, sections 4 and 5 provide the performance measure metrics and in detail comparative study of results.

2. Literature survey

Chauhan [1] produced a comparative analysis the most fundamental machine learning algorithms, such as logistic regression, decision trees, k-nearest neighbors, and naive bayes were used, and a maximum of 89% accuracy was obtained with the logistic regression. In many aspects, logistic regression is more user-friendly than its non-linear equivalent. This is especially true in terms of implementation, interpretation, and the ease with which it can be trained. It does not presume anything regarding the manner in which classes are laid out in the feature space. It uses a probabilistic method to make class predictions and is simple to apply in situations that include more than one class. The advantage and novel aspect is feature space distribution is used. When it comes to the feature space distribution of classes, it doesn't assume anything. It looks at class predictions made in view of probability and is easy to apply to situations with more than one class.

The authors Kuruvilla and Balaji [2] conducted research using a new approach. Using the Correlation-based feature selection with the multilayer perceptron technique, they were successful in predicting cardiac problems with an accuracy of 84.9057%. whereby CBFS and PCA are utilized in order to get the dimensionality down. When it comes to classification, a number of different algorithms, such as AdaBoost, Naïve Bays, MLP, and SMO are utilized. The CBFS-MLP hybrid achieves much better results than the MLP classifier does overall. When compared to the PCA-MLP combination, the CBFS-MLP model's accuracy metrics are more convincing. Rubini.P.E, Subasini, Katharine, [3] presented their work on random forest

Table 1. Literature review

Sr. No	Technique	Advantages	Disadvantages/Future Work
1	Hybrid decision support system [50]	A hybrid decision support system can be used in remote places when modern medical facilities are unavailable.	Only if a person has heart disease may it be diagnosed. This technique does not allow for the assessment of the degree of cardiac disease.
5	Machine Learning Algorithms Using Relief and LASSO Feature Selection Techniques [51]	results in a far better level of accuracy than comparable tasks.	The level of missing data influences the performance.
3	Cluster-based DT learning (CDTL) [52]	Attains high prediction accuracy.	The ideal decision tree's structure can be drastically altered by even the smallest change in the input.
2	Predicting heart disease risk using supervised learning and discrete weights [53]	Minimal false alarms, minimal process overhead, and maximum label prediction accuracy are all achieved with this system.	The dimensionality of different training corpus formats must be dealt with by employing ensemble classification procedures in the most efficient manner.
4	A technique based on global sensitivity analysis [54]	When picking attributes for classification, global sensitivity analysis is more important than individual feature selection approaches.	Requires high computation time for attribute selection.
6	Genetic algorithm (GA) with (RBF) radial basis function (GA-RBF) [55]	It also reduced the number of characteristics, which improved accuracy while also saving patients time and money.	The complex training process due to the large volume of data.
7	Ensemble Deep Learning and Feature Fusion [56]	To enhance heart disease prediction, low-dimensional and specialized weighted information must be extracted.	Data mining is required to improve the dataset for heart disease diagnostics.
8	Strength scores with significant predictors [57]	Achieved highest confidence score.	Accordingly, the machine learning approaches utilized in this research are confined to the most commonly used in heart disease prediction research.
9	Optimal ANN [58]	An appropriate method for analyzing large amounts of data in order to develop a heart disease prediction model.	The high computation time is the main drawback of this system.
10	Fuzzy rules are used in the Intelligent Big Data Analytics Model (IBDAM) [59].	This method has resulted in more accurate illness prediction.	The inaccurate data lead to lower accuracy.
10	Swarm Algorithms [41][47][48]	Developed swarm-based algorithm for Heart disease prediction	Accuracy needs to be improved

for the prediction of cardiac diseases. Their prediction accuracy was 84.81%. For the purpose of data categorization, they make use of a variety of methodologies, such as the NB, SVM, LR. Also, the R Forest method, which is part of the proposed model, is supposed to collect a lot of data about all of the factors and characteristics that cause coronary disease, train the data using the proposed artificial

Table 2. Dataset description table

Sr. No	Attributes	Description
1	Gender	0: male, 1: female
2	Age	Age of Patient in years
3	Education	Education of the person
4	C S	1: smoker 2: Non-smoker
5	CPD	cigarettes smoked daily by the individual
6	BP-Meds	Patients with BP Medication 0: not on BP Medication 1: on the Medication
7	PS	The patient previously suffered a stroke. 0: not having a previous stroke 1: previous stroke
8	P-Hyp	If the patient does not have hypertension, the score is 0. Furthermore, if hypertension is present, then the score is 1.
9	DE	Whether or not the patient is diabetic.
10	totChol	The amount of total cholesterol
11	sysBP	Systolic Blood Pressure
12	Dia-BP	Diastolic Blood Pressure
13	B-MI	Body Mass Index
14	HR	Heart Beats per/Minute
15	GS	Glucose-level
16	Ten Year CHD	Target- Risk of CHD fpr 10 years 0: No risk of CHD 1: Risk of CHD

intelligence (AI) computation, and then estimate the patient's likelihood of having a coronary disease.

Using random forest Masih and Ahuja [4] were also able to obtain an accuracy of 85.05% in the same year. This paper's strength is in its attempt to identify the best classifier for the Framingham dataset as well as other real-world data sets. It does so by providing an overview of the fundamentals of a number of well-known ML approaches and then applying those methods to the for mentioned datasets. Utilizing Random Forest led to the highest level of accuracy being achieved. Gaur [5] used 7 different machine learning approaches and the logistic regression showed 88% accuracy, the benefit of the model was it is split into 80:20 and trained. It made no assumptions about the feature space distribution of classes. It approached class predictions from a probabilistic standpoint and is simple to apply to scenarios involving more than one class.

The referral of nearby subject matter experts and medical facilities based on the client's preference would facilitate prompt and appropriate treatment. While medical care is a subject that is consistently expanding and creating a vast amount of data, there is a need to utilize the data for useful information, which attracts large organizations to invest heavily in this industry. Clinical errors could be decreased by integrating clinical decision support with computer-based patient records, increasing silent security, decreasing undesired practice variability, and enhancing learning outcomes. The program allows users to discuss heart-related topics. It then analyses client-specific information to determine if a certain disease is associated with it. Here, we will employ some intelligent data mining techniques to determine the ailment that is most likely to be diagnosed by the patient's nuances. The framework, therefore, displays the outcome-specific specialists for further therapy based on the outcome. The framework permits the consumer to view the specialist's information and may also be employed in the event of an emergency. Even if all of the earlier studies have employed a variety of methods and innovative methodologies, they have been unable to attain a better level of accuracy, which is something that is very necessary in the field of medicine.

3. Materials and methods

3.1 Materials

3.1.1 Description of dataset

The “Framingham” Dataset from Kaggle was used in this study since it is one of the most popular resources for researchers. The following table details its structure and contents: 4240 instances and 16 characteristics.

3.1.2. Data collection

Kaggle publishes a dataset on cardiac illness. This dataset comprises cardiovascular data from Framingham, Massachusetts. The usefulness of the suggested framework is evaluated by applying it to this dataset. The clustering is being done in order to categorize people into those who are more or less likely to acquire heart disease (HD) during the course of the following ten years. The collection includes not only clinical risk variables but also social risk factors and segment risk elements for patients. This collection is comprised of about 4,000 records and 16 attributes [40].

3.1.3. Data pre-processing

The data was first pre-processed, which includes data cleaning and handling of null and missing values. All discrepancies in the dataset were eliminated. There were no duplicate and missing values in the dataset.

3.1.4. Outlier detection and elimination:

Outlier identification [6] finds dataset observations that differ considerably from the norm and other values. Outliers. Eliminating outliers improves model accuracy. This study found outliers using box plots. Systolic and total cholesterol had outliers. Outliers in these columns were eliminated.

3.1.5. Balancing of data:

It was discovered that there were a strikingly disproportionate number of negative cases compared to positive cases. Since the dataset was extremely uneven as a result, it needs to be balanced in order to prevent issues from occurring while trying to fit the model. In this particular instance, the dataset was made more equitable by employing the SMOTE ENN technique. More members of the minority class were produced so that they would have an equivalent number to the members of the dominant class [7].

70% training “Framingham dataset” dataset is used, and 30% dataset for testing. The accuracy of various classifiers, such as RF, SVM, DT, KNN, ensemble techniques, and ANN, was calculated. It employed the Ensemble techniques of stacking, boosting, and bagging. The performance was enhanced using KNN.

3.2 Methods

3.2.1 Machine learning algorithms

SMOTE: SMOTE is a machine learning technique for handling challenges posed by dealing with sparse data. Since most algorithms may be easily crippled by improperly distributed data, we require these methods to improve the performance of our existing algorithms. In order to remedy the situation, balancing strategies were used prior to training the data in order to make it more even. It is a technique for improving data quality by creating simulated information from existing data [8]. SMOTE's advantage is that instead of creating duplicate data points, it generates synthetic data points that are slightly off from the originals. KNN is used to identify which neighboring nodes should be used when a minor is selected at random. With the random integer between 0 and 1, the current data point and the chosen neighbour’s vector get multiplied. The new data point is created by appending the resulting vector to the existing one. This method is analogous to moving the data point slightly in the direction of its neighbor. So, it can be made sure that your made-up observation is neither an exact copy of an existing observation nor too different from other known observations in your minority class.

SMOTE ENN: If we want to use the Edited Nearest Neighbor method, we must first identify the K-nearest neighboring observation and then check to see if the majority class of that neighbor matches the class of the present observation. If the neighbour’s majority class is different from the class of the observation, then both the neighbor and the observation are eliminated from the dataset. ENN employs a set of K=3 nearest neighbors in its default configuration. Similarly, it begins with a random subset of data from the minority class, calculates the average distance between that subset and its k nearest neighbors, multiplies that value by a random number between 0 and 1, and then adds that subset back to the minority class. This process keeps going until enough people from underrepresented groups have been added. In other words, SMOTE is now over, while ENN has begun. Count K, the number of your closest neighbors. Assume K=3 if it hasn't already been established. Seek for the K-nearest neighboring observation in the dataset, and then return the K-nearest neighbor's majority class. To clean up the data, observations and their K-nearest neighbors are discarded if their classes differ from those of the remainder of the dataset. Iterate until the required number of members from each group has been reached.

Logistic Regression: Using input variables, logistic regression models the likelihood of a discrete result. Logical regression excels at modelling two-valued outcomes [10] (true/false, yes/no, etc.). When there are more than two distinct discrete outcomes, multinomial logistic regression can model the situation. When trying to classify a fresh sample of data, logistic regression can help establish which group it best belongs in. In the field of cyber security, logistic regression is a useful analytical tool because many things, like detecting threats, can be seen as classification problems [11]. Log odds are related to an explanatory variable in the logistic regression model.

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i \tag{1}$$

Where $\log\left(\frac{p_i}{1-p_i}\right)$ is odds of function p_i , the explanatory variable is x_i, β_0, β_1 are the parameters of the model [12].

Naïve Bayes: It is a supervised ML method. It works by applying the Bayes theorem to a set of data and assuming that each piece of data is independent [13]. In other words, the method simply assumes that the input variables are all unrelated to one another [14].

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \quad (2)$$

A dependent feature vector is X and y is a class variable.

Decision Tree: Popular methods of machine learning include the decision tree algorithm. It's a scheme for classifying objects according to predetermined labels. Impurity is used to evaluate the consistency of a data set. In the case of a homogenous sample, all of the members belong to a single group [15]. Here are two of the several methods for gauging levels of impurity that are at your disposal.

I. Entropy
II. Gini impurity/index
III. Information Gain

I. The quantity of information required to fully describe a sample is its entropy. So, the entropy is 0 if all the elements in the sample are the same, and it is between 0 and 1 if the sample is spread out randomly.

$$\text{Entropy} = -\sum_{i=1}^{n} p_i \text{ x log } (p_i) \quad (3)$$

p_i = simply the frequentist probability of an element/class i in our data.

II. The Gini index refers to a measure of disparity in a statistical population. Values for it might be anything between zero and one. If the Gini index is 1, then the items in the sample are the most unequally distributed possible, whereas a value of 0 implies perfect equality. It is calculated by adding the squares of the probability associated with each group.

$$\text{Gini index} = 1\text{-}\sum_{i=1}^{n} p_i^2 \quad (4)$$

p_i = simply the frequentist probability of an element/class i in our data.

III. When data is partitioned according to an attribute, the information gain is proportional to the reduction in entropy. A decision tree may be built from the ground up by looking for the most informative attribute (i.e., the most homogeneous branches).

$$Gain(S,A) = Entropy(S) - \sum_{v\epsilon D_A} \frac{|S_v|}{|S|} Entropy(S_v) \quad (5)$$

| Sv | = summation of all node values
| S | = summation of set S node values
Entropy (Sv) = current node's entropy

There are various algorithms that are used to generate decision tree from data [16, 17].

Support vector machine (SVM): To be able to quickly classify new data points, the SVM algorithm finds the best lines (also called decision boundaries) dividing the n-dimensional space into classes. The hyperplane represents the decision boundary that is optimal in every way. The SVM selects the vectors and points that are the most extreme in order to construct the hyperplane. In most situations, support vectors are used to represent severe occurrences [19]. Mercer's theorem says that if the kernel matrix is Hermitian, positive, and semidefinite, then every pair of data points' kernel or Gram matrix assessment will also be positive and semidefinite. This means that

$$K(x,u)=\sum r\phi r(x)\ \phi r\ (u) \quad (6)$$

The Hilbert space is the place where $\phi(x)$ is defined. In other words, $\int\ \int K\ x(x,u)g(x)g(u)dxdu \geq 0\ \forall g(x)\int g2(x)dx<+\infty$

Various well-known kernel functions are

- When you require a linear kernel, use

$$K(x,u)=xT.u \quad (7)$$

- When you require a Polynomial Function, use

$$K(x,u)=(axT+c)q,q>0 \quad (8)$$

- When you require a Hyperbolic tangent(sigmoid), use

$$K(x,u)=\tanh(bxT+g) \quad (9)$$

- When you require a Gaussian radial basis function (RBF), use

$$K(x,u) = \exp\left(-\frac{||x-u||^2}{\sigma^2}\right) \quad (10)$$

- When you require a Laplacian radial basis function, use

$$K(x,u) = \exp\left(-\frac{||x-u||}{\sigma}\right) \quad (11)$$

- When you require a Randomized blocks analysis of variance (ANOVA RB) kernel, use

$$K(x,u) = \sum_{k=1}^{n} \exp\left(-\sigma(x^k - u^k)^2\right)^d \quad (12)$$

- When you require a Linear spline kernel in 1D, use

$$K(x,u)=1+x.u.\min(x,u)-\frac{x+u}{2}(\min(x,u)2+\frac{1}{3}\min(x,u)3)$$ [18, 19] (13)

K-nearest neighbors: It keeps a database of all the examples it has, and when classifying new examples, it gives similarity scores [22, 23]. The parameter K in KNN specifies how many nearest neighbors should be used in the majority-rule decision. With the help of the distance formula, the distance between the two locations will be calculated.

$$d2= ((x2-x1)^2+(y2-y1)^2) \quad (14)$$

The k values that are most near to the k factor will be chosen [22].

3.2.2. Ensemble techniques

Stacking: This ensemble approach combines various classifiers by using a meta-classifier. Multiple layers make up stacking, and each layer conveys its forecast to the layer above it [29]. The models in the layers underneath the top layer serve as the foundation for judgment. The dataset serves as the input for the lowest layer [30].

Bagging: A random selection of data from the initial dataset is chosen using an ensemble method known as bagging. It adds together the performances of each classifier to obtain the final result. Weak learners who have little bias and high variation are targeted by this approach. Bootstrapping, parallel training, and aggregation make up the three steps. Several subsets of data are initially assembled by replacing and picking data points at random. "Bootstrap replicates" are the name given to these datasets. Then, each of these data subsets is trained separately and simultaneously [31]. Following that, the results from each classifier are pooled using an average or a majority decision. One of the Bagging techniques is random forest.

Random Forest: To improve the accuracy of its predictions, several decision trees are used by Random Forest, and each is applied to different subsets of the input data and then averaging the results. The random forest utilises each decision tree's predictions rather than relying solely on one. and chooses the one with the most votes to determine the outcome [32, 18]. If you take the average of all the trees in a random forest, you'll get a sense of how important each individual trait is. When the weighted significance of every characteristic is totalled up for every tree and then split by the total tree count, then

$$RFfi_i = \frac{\sum_{j \epsilon all\, trees} norm\, fi_{ij}}{T} \quad (15)$$

$RFfi_i$ = values that are obtained from the Random Forest model's tree database to represent the significance of features

$norm\, fi_{ij}$ = the normalized significance of feature i in tree j

T = tally of trees.

Adaptive boosting: ADABOOST which is Adaptive Boosting, the name of a meta-algorithm for machine learning. To improve performance, it can occasionally be paired with a variety of other learning methods other learning algorithm outputs are combined to get a weighted total, which is the output of the boosted classifier (also known as "weak learners"). AdaBoost is adaptive in that it adjusts weak learners who succeed to favor samples of those earlier misclassified by classifiers. Even if each learner is bad, the final model can still be shown to converge to a good learner as long as their performance is slightly better than random guessing [33].

Gradient boosting: A Gradient boosting machine (GBM) takes the predictions from many decision trees and adds them together. It is important to know that all of the inefficient learners in a gradient-boosting machine are decision trees [34].

XG boosting: Extreme Gradient boosting, sometimes known as XG boost, is another popular boosting technique. The GBM algorithm has only been tweaked in the XG Boost technique. Operating similarly to GBM is XG Boost. In XG Boost, trees are built one after the other to fix problems with the trees that came before [33].

Cat boost: As the name suggests, Cat boost is an effective boosting technique for data that contains categorical variables. Most machine learning algorithms are incapable of dealing with input that includes strings or categories. Because of this, making category variables into numbers is an important part of pre-processing. It can handle different categories of information internally [36]. Many kinds of statistics on feature combinations are used to convert them to numbers.

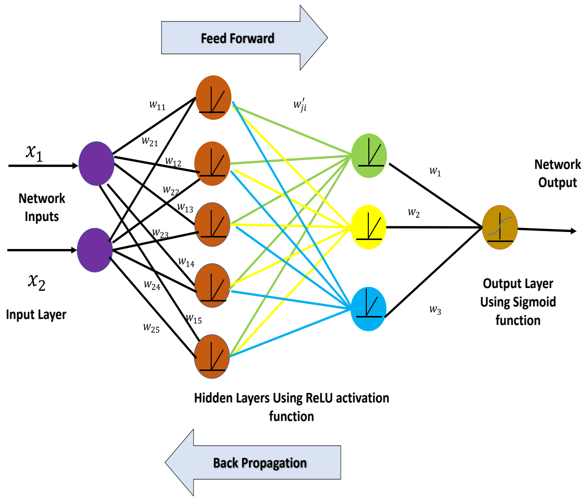

Figure. 1 One input layer, two hidden layers, and an output layer comprise an artificial neural network

3.2.3. Artificial neural network(ANN):

An ANN is a type of computer model that consists of a group of processing modules that take in data and make decisions based on a predetermined set of guidelines [25, 45]. These modules are joined in a network that resembles the structure of genuine neural networks and is designed to emulate their behavior. In addition to the input and output layers, an artificial neural network may also include a layer known as a "hidden layer," which is situated in the middle of the two layers. After receiving a number of weighted inputs from the layer below them, the artificial neurons in this layer of the network make use of something called an "activation function" in order to generate outputs.

The connection between the neuron's inputs and outputs is shown by the weighted, directed edges in Fig. 1. The artificial neural network receives its input as features of the dataset’s vector picture. Then, every nth input (n) is given one of the feature values using the notation x(n) [26]. Then, the weights are applied to each input individually. These weights are used to represent the quality of connections between neurons in an artificial neural network. [27] If the weighted sum is zero, the output is rendered non-zero by adding bias. Here, The activation function receives the sum of the weighted inputs, and a maximum value is used to keep the response within the target range [28].

4. Proposed model

4.1 KNN

K-nearest neighbors (KNN), a supervised learning technique, is used for both classification and regression. By calculating the distance between each training point and the test data, KNN attempts to predict the appropriate class for the input data. Then, choose the K number of points that most closely resemble the test data. After determining how likely it is that the test data will fall into each of the "K" training data classes, the KNN approach selects the class with the highest probability. In regression, the value is the mean of the 'K' training points selected. The KNN algorithm can compete with the most precise models because its predictions are so precise. Hence, Applications that demand great accuracy but don't need a model that can be read by humans can employ the KNN approach. The accuracy of the forecasts is impacted by the distance measurement. The KNN method is therefore useful for situations for which adequate domain knowledge is available. This understanding facilitates the selection of a suitable measure. The KNN approach is a type of idle learning in which categorization is done before prediction-generating calculations are done. Even though this method costs more to calculate than other ones, it is still the best choice in some circumstances where accuracy is more critical than the frequency of prediction requests.

KNN keeps a database of all the examples it has, and when classifying new examples, it gives similarity scores [22, 23]. The parameter K in KNN specifies how many nearest neighbors should be used in the majority-rule decision. With the help of the distance formula, the distance between the two locations will be calculated.

$$d2= ((x2-x1)\ ^2+(y2-y1)\ ^2) \qquad (16)$$

The k values that are most near to the k factor will be chosen [24].

4.2 Architecture and working:

It was crucial to incorporate all of the necessary information in the dataset before getting started with the analysis. The Framingham data collection was utilised in order to meet the aims of this study. The data's outliers were identified and then modifications were made to account for them. The dataset that was used for this inquiry had some data that was imbalanced. A strategy that is SMOTE and SMOTE ENN was applied to attain this purpose. The purpose of this technique was to ensure that the dataset was correct and that underrepresented groups were given the representation that they deserved. A number of different classifiers, including SVM, DT, KNN, and Naive Bayes, as well as a number of different ensemble learning methods, including stacking,

Received: March 21, 2023. Revised: July 18, 2023.

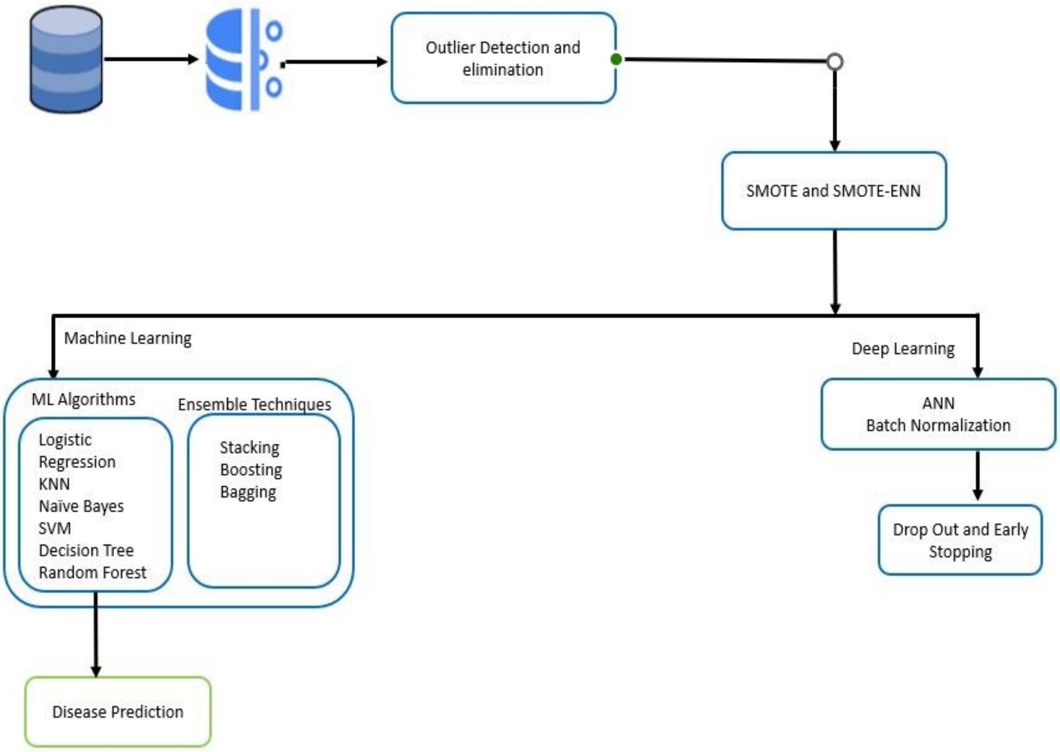

Figure. 2 Data flow diagram of proposed model

boosting, and bagging, random forests, have been implemented in order to improve the accuracy of the risk assessment for coronary heart disease ANNs with an Adam optimizer, a sigmoid activation function for the output layer, and a ReLu activation function for the hidden layers were utilised to train the neural network. The methods are analyzed and contrasted with their relative effectiveness. Then, in the deep learning strategy that was proposed for ANN, The hidden layer is activated with the ReLu activation function in conjunction with the Adam optimizer, while the final layer is activated with a sigmoid activation function. It calculated the epoch values and the accuracy by using graphs of loss and accuracy. Next, dropout layers and batch normalization are used to identify the validated accuracy and validated loss. Finally, it is found that the validated loss and validated accuracy can be used using dropout layers. All are judged on their performance or accuracy. The following Fig. 2 is a description of the proposed system's architecture.

5. Results and analysis

Experiments on the dataset using a number of well-known algorithms were carried out in order to compare the effectiveness of our method to other machine learning techniques and assess its performance. Python was the platform that was used for the testing.

5.1 Experiment setup

The Jupyter Notebook is used to conduct an analysis of the performance of the ML models in this section. Users have the option to put all the parts of a data project together in one place, which makes it much easier to explain the project's entire process to the audience you have in mind. It offers a library of different models that may be used for things like data preparation, classification, clustering, forecasting, visualization, and so on. The personal computer (PC) that served as the testing environment possessed each of the following qualities: An x64-based CPU with an 8 MB cache, 4 cores, 8 threads, and a frequency range of 2.40 GHz to 4.20 GHz, as well as Windows 11, a 64-bit operating system, is an x64-based processor.

5.2 Performance metrics

Experiments were conducted on the dataset using a variety of well-known algorithms to see how well our method fared in contrast to other machine learning techniques. Python was used as the testing platform. Classification models forecast data sample targets in classification tasks. Thus, the categorization model predicts event class probabilities. To solve actual problems reliably, the

Table 3 Understanding of confusion matrix

Prediction	Actual Value	Type	Explanation
1	1	True Positive	Predicted Positive and was Positive
0	0	True Negative	Predicted Negative and was Negative
1	0	False Positive	Predicted Positive and was Negative
0	1	False Negative	Predicted Negative and was Positive

Table 4 Confusion matrix

	Predicted Class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

categorization model must be tested before use. Key performance indicators can assess a machine learning classification model. F1-score, recall, and accuracy are examples. Machine learning relies on precise prediction; therefore, understanding a model's performance is crucial. Underpinning accuracy, recall, and true positives, true negatives, false positives, and false negatives are all parts of the F1-Score. They underpin accuracy, recall, and F1-Score [37].

"True positive" (TP) signifies that both the actual and anticipated classes have value. The positive value was anticipated properly as shown in Table 3. True negatives (TN) are real-class projected negative values. When the observed class differs from the planned class, false positives and negatives happen. False positives occur when the projected class is correct but the actual class is incorrect (FP). When the actual class differs from the expected class, false negatives happen (FN) as shown in Table 4 [38].

5.2.1. Performance parameters

Accuracy: Accuracy is the easiest performance measure to understand. Our model suggests it's best. Accuracy requires near-true positive and false negative rates in a dataset. Thus, while assessing your model's efficacy, you must consider more factors [39].

$$Accuracy = \frac{TN+TP}{TN+FP+TP+FN} \tag{17}$$

Precision: It is the ratio of genuine positives to positive samples.

$$Precision = \frac{TP}{TP+FP} \tag{18}$$

TP + FP (denominator) > TP (Numerator) reduces machine learning model accuracy (Numerator). TP (Numerator) > TP + FP (denominator) indicates good machine learning model accuracy (denominator). Thus, precision helps us assess the machine learning model's dependability and determine if it's favorable.

Recall: To calculate recall, divide the number of positive samples by the quantity accurately categorized as positive. Recall measures the model's positive sample detection. Recall increases positive sample detection.

$$Recall = \frac{TP}{TP+FN} \tag{19}$$

TP + FN (denominator) > TP (Numerator) reduces machine learning model recall (Numerator). ML model recall is strong when TP (numerator) > TP + FN (denominator) (denominator). The recall is independent of sample categorization errors, whereas accuracy is not. If the model finds all positive samples to be positive, the recall will be 1 [40].

F1-score: A binary classification model's F1 score indicates accuracy. It's computed accurately. A particular score that considers accuracy and memory. The F1 score can be found by using a harmonic mean to give equal weight to both accuracy and recall.

$$F1 - score = 2 * \frac{precision * recall}{precision + recall} \tag{20}$$

The F1 score should be used when accuracy or recall are more important than the other because it includes both. Avoiding false negatives may be more important than false positives.

5.3 Analysis

To ensure the efficacy of the suggested method, it is compared to well-known machine learning techniques. Logistic regression, SVM, decision trees, KNN, Naive Bayes, random forest, Bagging Meta estimator, Gradient boosting, AdaBoost, cat boost, XG boost, light GBM, and stacking are among the methods that may be used. In the table below, the results of the tests done on the Framingham datasets using each of the different methods are summarized. Table 5 and Fig. 3 demonstrate that while utilizing the model pipeline, the accuracy achieved by KNN is the best (97.85%), while that achieved by logistic regression is the lowest (76.88%) The accuracy of the decision tree classifier is 85.38%, which is higher than the accuracy of the support vector machine method (79.46%).

Table 5 ML algorithms using model pipeline

S. No.	Method	Accuracy Score (%)
1	Logistic Regression	76.88%
2	SVM	79.46%
3	Decision Tree Classifier	85.38%
4	K Neighbors Classifier	97.85%

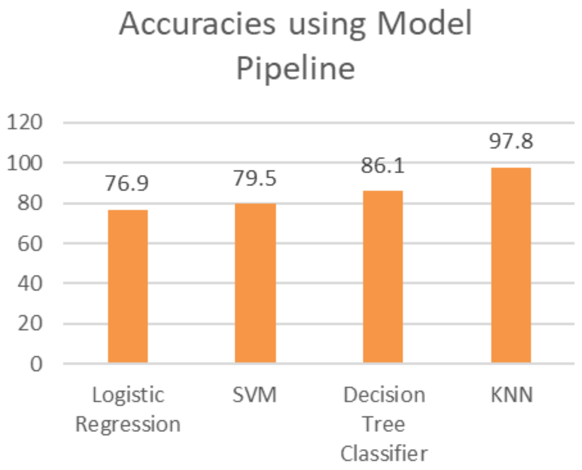

Figure. 3 Accuracy graph for ML algorithms

Table 6. Accuracy table without grid SearchCV method

S. No.	Method	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
1	Logistic Regression	76.9	78.8	86.4	82.399
2	SVM	79.5	82.6	85.2	83.899
3	Decision Tree Classifier	86.1	87.2	91.3	89.2
4	KNN	97.8	97.5	99.1	98.3
5	Naïve Bayes	73.9	78.9	79.80	79.4

Table 7 Accuracy of grid search method

Method	Accuracy Score (%)	Precision Score (%)	Recall Score (%)	F1 Score (%)
Logistic Regression	77.5	77	78	77
SVM	82.89	83	83	83
Decision Tree Classifier	82.89	83	83	83
KNN	96.83	98	98	98

Table 8 Accuracy of ensemble techniques

S. No.	Method	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
1	Bagging meta-estimator Random Forest	89.8 92.5	90 93	90 92	90 92
2	Boosting Ada Boost Classifier Gradient Boosting XG Boosting LightGBM Cat Boost	68.49 74.62 74.62 90.10 88.70	91 74 82 90 89	91 75 82 90 89	91 74 82 90 89
3	Stacking	94.30	94	94	94

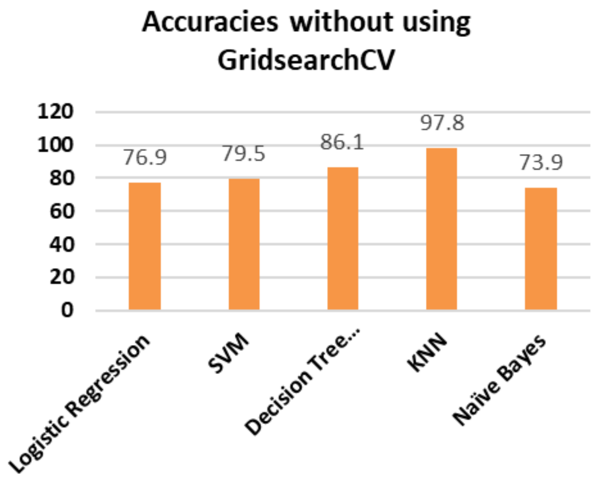

Figure. 4 Accuracy graph without hyperparameter tuning

When we take a look at Fig. 4 and Table 6, we can see that KNN obtains an accuracy of 97.8% without having to resort to grid search, but the most that Decision Tree can do is acquire an accuracy of 86.1%. When compared to logistic regression and the Naive Bayes classifier (76.9%), support vector machines (SVM) only attain an accuracy of 79.5%.

When we examine the data in Table 7 and Fig. 5, we can see that the use of gridsearchCV helped KNN achieve the maximum level of accuracy. This approach exceeds both SVM and decision trees, which are only able to achieve an accuracy of 82.89%, by achieving a level of accuracy that is 96.83% more accurate than this method. On the other hand, the accuracy of the logistic regression comes in at 77.5%. With an accuracy of 94.30%, stacking surpasses the other approaches presented in Table 8 and Fig. 6, which are then followed by random forest (92.5%)

Received: March 21, 2023. Revised: July 18, 2023.

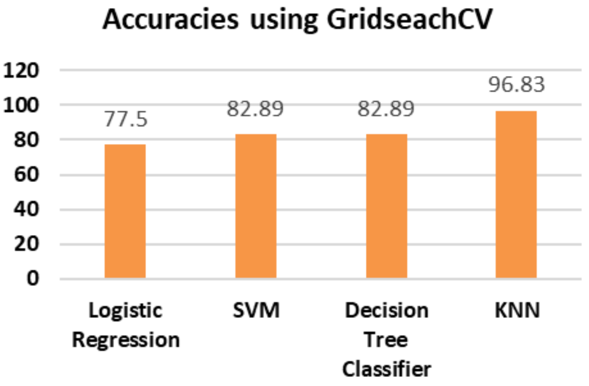

Figure. 5 Accuracy graph with hyperparameter tuning

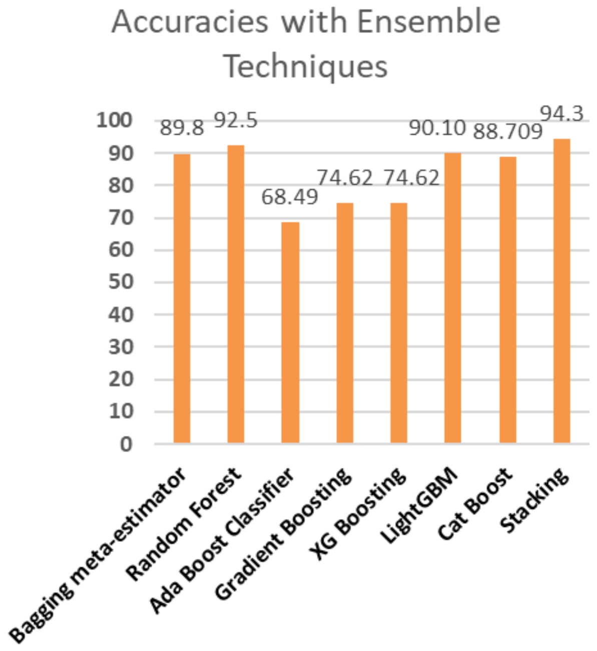

Figure. 6 Accuracy graph using ensemble technique

Table -9 ANN accuracy

S. No.	Method	Accuracy (%)	Classifier details
1	ANN	84.5	Optimizer = 'Adam' Activation function for Hidden layer = 'Relu' Output layer = 'Sigmoid'

and light GBM (90.107%). The bagging meta-estimator is currently at 89.8%, while the cat boost meta-estimator is currently at 88.7%. Both the gradient and the XG boost attain comparable levels of accuracy (74.62%) when measured against Ada's boost, which is 68.49%.

An accuracy of 84.5% was achieved by using the results of employing the Adam optimizer and Relu activation for the hidden layers combined with Sigmoid activation for the output layers in ANN from Table 9.

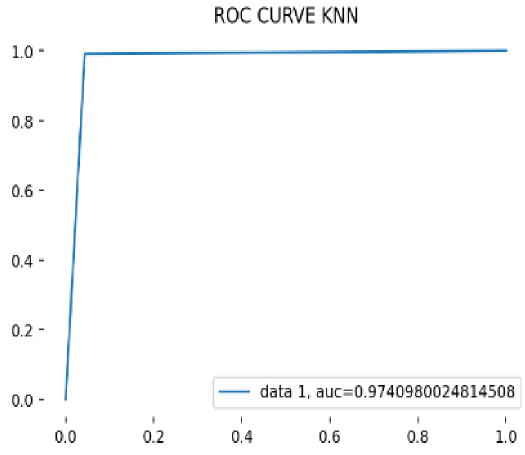

Figure. 7 AUC ROC curve

Table 10 Comparison table

Methods	Accuracies (%)	Accuracies of [1] (%)
Logistic regression	76.9	89
KNN	97.8	88
SVM	79.5	88
DT	86.1	80
Random Forest	92.5	87
ROC	97.4	71.933

The results of evaluating the system for recognizing heart disease are depicted as an area under the receiver operating characteristic curve (AUC). The interpretation of the AUC curve is presented in Figure 7, with the assistance of test results taken from the Framingham dataset. In our case, the AUC value is 0.974, which is very close to being equal to 1, indicating that the performance to detect heart diseases has been greatly improved.

6. Case study

In the previous section, various contrasts and comparisons were drawn between the recently presented approach and a number of other algorithms that are commonly used. The findings of the experiments provide credence to the assertion that the recommended strategy is preferable. A comparative study was undertaken with previously published techniques that made use of the Framingham heart disease dataset. The purpose of this investigation was to establish that the recommended method is superior to the methods in question. The outcomes of these comparisons are presented in Table 1, which can be found below. Here, it can be seen that our method offers a substantial benefit. Table 10 demonstrates that there has been an increase in accuracy, which is encouraging news about the prediction of diseases. The use of smote and The ENN has increased the accuracy of k-nearest neighbour, decision tree, and random forest but decreased the accuracy of logistic regression and support vector machine.

7. Conclusion and future work

Heart disease is any ailment that hinders the functioning of the heart. Nowadays, the major cause of death is heart disease. Cigarettes, alcohol, a diet heavy in fat, and a sedentary lifestyle all raise the risk of cardiovascular disease. According to the WHO, nearly 10 million people die annually from heart disease. To prevent heart disease, a healthy lifestyle and early diagnosis are required. Diagnostics and treatment are the primary focuses of contemporary healthcare. Worldwide, heart disease is the leading avoidable cause of death. Disease identification determines sickness treatment. The suggested method diagnoses heart abnormalities at an early stage to prevent adverse consequences. The use of machine learning in healthcare is now garnering a great deal of interest owing to its potential to improve both knowledge and, therefore, patient outcomes. Due to its ability to improve both knowledge and, therefore, patient outcomes machine learning has aided in the accurate detection of a variety of ailments. The precise diagnosis of a wide range of diseases has been facilitated by ML. Future research assisted by many and strong machine learning computations has the potential to provide precise infection progression forecasting and patient treatment. Daily, the healthcare industry generates a deluge of data pertaining to medical services. This data may be mined for information on disease trends and treatment outcomes. The medical records of the patient include private information that will, at some point in the future, be used to give the patient the appearance that they are more involved in decisions about their health. Similarly, this area necessitates expansion in the form of educational data use in the medical profession.

With an emphasis on parameter tuning, ensemble methodologies, and recursive feature elimination procedures, the proposed model is designed with the expectation that its primary role will be to improve the accuracy and reliability of cardiac illness prediction. Our prediction techniques included logistic regression, decision trees, K-nearest neighbor (KNN), support vector machine (SVM), and Naive Bayes machine learning approaches, ensemble technique approaches, and artificial neural network (ANN) with a focus on regularization. By using the provided paradigm to the dataset, its utility is assessed. Clustering is performed in order to categorise individuals according to their likelihood of developing heart disease (HD) during the next 10 years. In addition to clinical risk data, the collection covers societal risk factors and segment risk aspects for patients. This collection has around 4,000 records and sixteen qualities. This study examined the dataset to assess whether or not it was well-balanced. It was revealed that the number of negative instances was remarkably disproportionate to the number of positive ones. Due to the fact that the dataset was severely unbalanced, it must be rebalanced in order to avoid problems while attempting to fit the model. In this case, the dataset was made more equal by the use of SMOTE ENN. More members of the minority class were generated so that their numbers would be comparable to those of the majority class. Before beginning the analysis, it was essential to include all of the relevant information in the dataset. This research used the Framingham data collection in order to achieve its objectives. The data outliers were identified and then adjustments were made to account for them. The dataset utilised for this investigation had some unbalanced data. A SMOTE and SMOTE ENN method was used to achieve this objective. Experiments were conducted on the dataset using a variety of well-known algorithms to compare the performance of our technique to that of other machine learning tactics. The suggested approach was intended to improve the resilience, reliability, and accuracy of coronary heart disease risk estimates. In this study, both deep learning and machine learning were used. To improve accuracy, it may use recursive feature reduction to isolate the most essential properties. KNN demonstrated to be the most accurate ensemble approach when compared to stacking, boosting, and bagging. In prior experiments, the proposed model attained the greatest levels of accuracy (97.8%). The findings are positive, with increased performance prediction in comparison to more conventional methods. Relative to other algorithms, some ones provide superior performance. The project's purpose is to identify classifiers capable of predicting cardiac disease with adequate accuracy for practical use. This research retrieves the data set from the Kaggle website. Using a range of statistical models, such as Naive Bayes, decision trees, LR, SVM, and KNN, the early prediction of heart illness is examined. On the basis of patient characteristics and data, ensemble methods including as bagging, boosting, stacking, and random forest have been used to predict cardiovascular disease. In this experiment, KNN contributed to greater accuracy.

Future research aims to implement machine learning's categorization algorithms to accurately anticipate cardiac disorders. In addition, the model might be improved by including more data and techniques. The long-term objective of this article is to be able to forecast cardiac disease utilising modern methodologies, algorithms, and a high degree of precision in less time.

Conflict of interest

All the authors declare that "There are no conflicts of interest".

Author contributions

This research has been conceptualized and methdology implemented by Snehal Bankatrao Shinde, Kankipati Lahari, Keerthika Chowdary, Garimella and Vicharapu Sowmya Sree. Analysis and investigation of research is done by Nileshchandra K Pikle and Girish Bhavekar. Research article review and editing is done by Pradnya Borkar and Sagarkumar Badhiye.

References:

[1] Y. Chauhan, "Cardiovascular Disease Prediction using Classification Algorithms of Machine Learning", *International Journal of Science and Research (IJSR)*, Vol. 9, No. 5, pp. 194-200, 2020.

[2] A. Kuruvilla and N. Balaji, "Heart disease prediction system using Correlation Based Feature Selection with Multilayer Perceptron approach", In: *Proc. of IOP Conference Series: Materials Science and Engineering*, pp. 1-6, 2021.

[3] P. Rubini, C. Subasini, A. Katharine, V. Kumaresan, S. Gowdham, and T. Nithya, "A Cardiovascular Disease Prediction using Machine Learning Algorithms", *Annals of the Romanian Society for Cell Biology*, Vol. 25 No. 2, pp. 904–912, 2021.

[4] N. Masih and S. Ahuja, "Prediction of Heart Diseases Using Data Mining Techniques: Application on Framingham Heart Study", *International Journal of Big Data and Analytics in Healthcare*, Vol. 3, No. 2, pp. 1-9, 2018.

[5] H. Jindal, S. Agrawal, R. Khera, R, Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms", In: *Proc. of IOP Conference Series: Materials Science and Engineering*, Vol. 1022, No. 1, 2021.

[6] M. Batchanaboyina, and N. Devarakonda, "Efficient Outlier Detection for High Dimensional Data using Improved Monarch Butterfly Optimization and Mutual Nearest Neighbors Algorithm: IMBO-MNN", *International Journal of Intelligent Engineering and Systems*, Vol. 13, No. 2, pp. 63-73, 2020, doi: 10.22266/ijies2020.0430.07.

[7] M. Raju and B. Rao, "Colorectal Cancer Disease Classification and Segmentation Using a Novel Deep Learning Approach", *International Journal of Intelligent Engineering and Systems*, Vol. 15, No. 4, pp. 227-236, 2022, doi: 10.22266/ijies2022.0831.21.

[8] B. Darst, K. Malecki, and C. Engelman, "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data", *BMC Genet*, Vol. 19, No, 65, pp. 1-6, 2018.

[9] P. Theerthagiri and P. Venkatesan, "Cardiovascular Disease Prediction using Recursive Feature Elimination and Gradient Boosting Classification Techniques", *arXiv preprint arXiv:2106.08889*, 2021.

[10] U. Dulhare, "Prediction system for heart disease using Naive Bayes and particle swarm optimization", *Biomedical Research*, Vol. 29, No. 12 pp. 2646-2649, 2018.

[11] J. Hosmer, W. David, S. Lemeshow, and R. Sturdivant, "Introduction to the Logistic Regression Model", *John Wiley & Sons*, Vol. 398, 2013.

[12] M. Saw, T. Saxena, S. Kaithwas, R. Yadav and N. Lal, "Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning", In: *Proc. of International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, pp. 1-6, 2020.

[13] C. Shenglei, I. Geoffrey, L. Liu, and M. Xin, "A novel selective naïve Bayes algorithm", *Knowledge-Based Systems*, Vol. 192, 2020.

[14] H. Zhang, "The Optimality of Naive Bayes", In: *Proc. of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, FLAIRS, 2004.

[15] Y. Song and L. Ying, "Decision tree methods: applications for classification and prediction.", *Shanghai Archives of Psychiatry*, Vol. 27, No. 2, pp. 130-135, 2015.

[16] S. Rasoul and D. Landgrebe, "A survey of decision tree classifier methodology", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 21, No. 3, pp. 660-674, 1991.

[17] S. Maji and S. Arora, "Decision Tree Algorithms for Prediction of Heart Disease", In: Fong, S., Akashe, S., Mahalle, P. (eds), "Information and

Communication Technology for Competitive Strategies", *Lecture Notes in Networks and Systems*, Vol. 40, 2019.

[18] S. Wager, "Comments on: A random forest guided tour", *Test*, Vol. 25, No. 2, pp. 261-263, 2016.

[19] M. Chandra and S. Bedi, "Survey on SVM and their application in image classification", *International Journal of Information Technology*, Vol. 13, No. 5, pp. 1-11, 2021.

[20] R. Kumar, S. Srivastava, A. Dass, Anuli, and S. Srivastava, "A novel approach to predict stock market price using radial basis function network", *International Journal of Information Technology*, Vol. 13, No. 6, pp. 2277-2285, 2021.

[21] N. Lutimath, C. Chethan, and S. Basavaraj, "Prediction of heart disease using machine learning", *International Journal of Recent Technology and Engineering*, Vol. 8, No. 2 pp. 474-477, 2019.

[22] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN Model-Based Approach in Classification", In: *Meersman, R., Tari, Z., Schmidt, D.C. (eds) on The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003. Lecture Notes in Computer Science*, Vol. 2888, 2003.

[23] N. Khateeb and M. Usman, "Efficient heart disease prediction system using K-nearest neighbor classification technique", In: *Proc. of the International Conference on Big Data and Internet of Thing*, 2017.

[24] L. Sharma, H. Chhabra, and U. Chauhan, "Mental arithmetic task load recognition using EEG signal and Bayesian optimized K-nearest neighbour", *International Journal of Information Technology*, Vol. 13, No. 6, pp. 2363-2369, 2021.

[25] M. Mishra and M. Srivastava, "A view of Artificial Neural Network", In: *Proc. of International Conference on Advances in Engineering & Technology Research*, Unnao, India, pp. 1-3, 2014.

[26] J. Zou, Y. Han, and S. So, "Overview of Artificial Neural Networks", In: *Livingstone, D.J. (eds) Artificial Neural Networks. Methods in Molecular Biology™*, Humana Press, Vol. 458, 2008.

[27] J. Bailer, A. Coryn, R. Gupta, and H. Singh, "An introduction to artificial neural networks", *arXiv Preprint*, 2001.

[28] J. Kim and S. Kang, "Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis", *J. Healthc. Eng*, 2017.

[29] D. Mane, R. Tapdiya, S. Shinde, "Handwritten Marathi numeral recognition using stacked ensemble neural network", *International Journal of Information Technology*, Vol. 13, No. 5, 2021.

[30] R. Vasudev, B. Anitha, G. Manikandan, B. Karthikeyan, L. Ravi, and V. Subramaniyaswamy, "Heart disease prediction using stacked ensemble technique", *Journal of Intelligent & Fuzzy Systems*, Vol. 39, No. 6, pp. 8249-8257, 2020.

[31] L. Breiman, "Bagging predictors", *Machine Learning*, Vol. 24, pp. 123–140, 1996.

[32] M. Tu, D. Shin and D. Shin, "Effective Diagnosis of Heart Disease through Bagging Approach", In: *Proc. of II International Conference on Biomedical Engineering and Informatics, Tianjin*, China, pp. 1-4, 2009.

[33] R. Schapire, D. Denison, M. Hansen, C. Holmes, B. Mallick, and B. Yu, "The Boosting Approach to Machine Learning: An Overview", In: *Nonlinear Estimation and Classification. Lecture Notes in Statistics*, Springer, New York, Vol. 171, 2003.

[34] N. Alexey and K. Alois, "Gradient boosting machines, A Tutorial", *Frontiers in Neurorobotics*, Vol. 7, 2013.

[35] T. Chen and G. Carlos, "Xgboost: A scalable tree boosting system", In: *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[36] A. Ibrahim, R. Ridwan, M. Muhammed, R. Abdulaziz, and G. Saheed, "Comparison of the CatBoost classifier with other machine learning methods", *International Journal of Advanced Computer Science and Applications*, Vol. 11, No. 11, pp. 738-748, 2020.

[37] P. Flach, "Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward", In: *Proc. of the AAAI Conference on Artificial Intelligence*, Vol. 33, No. 01, pp. 9808-9814, 2019.

[38] A. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease", *Neural Computing and Applications*, Vol. 29, No. 10, pp. 685-693, 2019.

[39] A. Patil and S. Subbaraman, "Performance analysis of static hand gesture recognition approaches using artificial neural network, support vector machine and two stream based transfer learning approach", *International*

Journal of Information Technology, Vol. 14, pp. 3781-3792, 2022.

[40] M. Buckland and F. Gey. “The relationship between recall and precision”, *Journal of the American society for Information Science*, Vol. 45, No. 1, pp. 12-19, 1994.

[41] A. Goswami, G. Bhavekar, and P. Chafle, “Electrocardiogram signal classification using VGGNet: a Neural Network based classification model”, *International Journal of Information Technology*, Vol. 15, pp. 119–128, 2022.

[42] https://www.kaggle.com/code/lauriandwu/machine-learning-heart-disease-framingham

[43] A. Bojamma and C. Shastry, “A study on the machine learning techniques for automated plant species identification: current trends and challenges”, *International Journal of Information Technology*, Vol. 13, pp. 989–995, 2021.

[44] M. Divate, “Sentiment analysis of Marathi news using LSTM”, *International Journal of Information Technology*, Vol. 13, pp. 2069–2074, 2021.

[45] T. Kulkarni and N. Dushyanth, “Performance evaluation of deep learning models in detection of different types of arrhythmia using photo plethysmography signals”, *International Journal of Information Technology*, Vol. 13, No. 6, pp. 2209-2214, 2021.

[46] N. Pandey and N. Muppalaneni, “A novel algorithmic approach of open eye analysis for drowsiness detection”, *International Journal of Information Technology*, Vol. 13, pp. 2199-2208, 2021.

[47] G. Bhavekar and A. Goswami, “A hybrid model for heart disease prediction using recurrent neural network and long short term memory”, *International Journal of Information Technology*, Vol. 14, pp. 1781–1789, 2022.

[48] G. Bhavekar and A. Goswami, “Herding Exploring Algorithm with Light Gradient Boosting Machine Classifier for Effective Prediction of Heart Diseases”, *International Journal of Swarm Intelligence Research (IJSIR)*, Vol. 13, No. 1, pp. 1-22, 2022.

[49] C. Vasantrao and N. Gupta, “Wader hunt optimization based UNET model for change detection in satellite images”, *International Journal of Information Technology*, pp. 1-13, 2023.

[50] P. Rani, R. Kumar, N. Ahmed, and A. Jain, “A decision support system for heart disease prediction based upon Machine Learning”, *Journal of Reliable Intelligent Environments*, Vol. 7, pp. 263–275, 2021.

[51] V. Esposti, F. Fuschini, H. Bertoni, R. Thoma, T. Kurner, X. Yin, and K. Guan, “IEEE Access Special Section Editorial: Millimeter-Wave and Terahertz Propagation, Channel Modeling, and Applications”, *IEEE Access*, Vol. 9, pp. 67660-67666, 2021.

[52] G. Magesh and P. Swarnalatha, “Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction”, *Evolutionary Intelligence*, Vol. 14, pp. 583–593, 2021.

[53] F. A. Yarimi, N. Munassar, M. Bamashmos, and M. Ali, “Feature optimization by discrete weights for heart disease prediction using supervised learning”, *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, Vol. 25, No. 3, pp. 1821-1831, 2021.

[54] G. Saranya A. Pravin, “Hybrid Global Sensitivity Analysis Based Optimal Attribute Selection Using Classification Techniques by Machine Learning Algorithm”, *Wireless Personal Communications*, Vol. 127, pp. 2305–2324, 2022.

[55] B. Doppala, D. Bhattacharyya, M. Chakkravarthy, and T. Kim, “A hybrid machine learning approach to identify coronary diseases using feature selection mechanism on heart disease dataset”, *Distributed and Parallel Databases*, Vol. 41, pp. 1–20, 2023.

[56] F. Ali, S. E. Sappagh, S. Islam, D. Kwak, A. Ali, M. Imran, and K. Kwak, “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.

[57] A. Yazdani, K. Varathan, Y. Chiam, A. Malik, and A. Ahmad, “A novel approach for heart disease prediction using strength scores with significant predictors”, *BMC Medical Informatics and Decision Making*, Vol. 21, No. 194, 2021.

[58] R. Selvi and I. Muthulakshmi, “An optimal artificial neural network based big data application for heart disease diagnosis and classification model”, *Journal of Ambient Intelligence and Humanized Computing*, Vol. 12, pp. 6129–6139, 2021.

[59] M. Pandian, “Intelligent Big Data Analytics Model for Efficient Cardiac Disease Prediction with IoT Devices in WSN using Fuzzy Rules”, *Wireless Personal Communications*, Vol. 127, 2021.