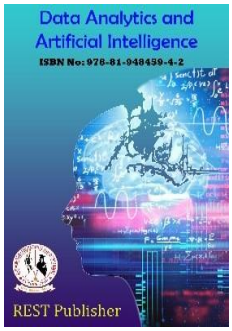
Data Analytics and Artificial Intelligence
Vol: 3(6), 2023
REST Publisher; ISBN: 978-81-948459-4-2
Website: http://restpublisher.com/book-series/daai/



Enhancing the Classification Techniques for Heart Disease Prediction

[*]P Sharath Prabhu, Dr. T. Jemima Jebaseeli
KarunyaInstitute of Technology and Sciences, Coimbattore, Tamil Nadu, India
*Corresponding Author Email:psharathprabhu@karunya.edu.in

Abstract:*Heart disease detection is important to prevent the sudden casualties for an individual. Machine learning algorithms have shown promising results in predicting heart disease, but their performance varies depending on the dataset and algorithm used. In this study, we compare the performance of seven popular algorithms: Support Vector Machine (SVM), Decision Tree, Logistic Regression, XG-Boost, k-Nearest Neighbours (KNN), Naive Bayes, and Random Forest.The dataset used in this study contains parameters such as age, sex, blood pressure, and cholesterol level. We pre-processed the data by handling missing values and normalizing the numerical attributes. Then, we trained and evaluated the algorithms using fold cross-validations for hyper Tuning and measured their performance based on accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC). The accuracy attained with Logistic Regression, SVM, Decision Tree, KNN, XG-Boost, Naive Bayes and Random Forest is compared for various data sets. In conclusion, we highlight the potential of fine-tuned algorithm in predicting heart disease that provides insight into the performance of different classification techniques [1]. Hence an efficient model is developed for early detection and prevention of heart disease.*

Keywords:*cardiovascular disease Prediction, Classification of Algorithms, Hyper-parameter tuning, Random search, model prediction*

1. INTRODUCTION

Heart disease continues to be a primary cause of death and is a major global health issue. Effective prevention and treatment of heart disease depend on early diagnosis and effective prediction of the condition. Machine learning techniques have gained more and more attention in recent years due to it's potential to correctly forecast the development of cardiac disease. In this paper, we explore the use of SVM, decision tree, logistic regression, XG-Boost, KNN, Naive Bayes, and Random Forest for heart disease prediction.Among these seven Algorithms, there are three algorithms which is hypertuned and whose accuracy is greater than the past papers. Machine learning techniques, such as XGBoost, have shown promise in predicting heart disease. The powerful gradient boosting algorithm XGBoost has proven to perform better than others in a variety of applications, including medical diagnosis. In this paper, we explore the potential of XGBoost in heart disease prediction and evaluate its performance using relevant clinical data.Logistic regression models can be trained on large datasets of patient information to identify the likelihood of heart disease with given parameters. Health care providers can identify high-risk patients and offer individualised care to prevent or control heart disease by utilising the power of machine learning.KNN is an effective method for classification and prediction tasks. In order to detect and treat heart disease quickly, medical professionals can use KNN to analyse patient data and forecast the possibility of heart disease.In this paper, we will explore the use of KNN in heart disease prediction and its potential benefits for patient health.

2. LITERATURE REVIEW

Logistic regression has been widely used in predicting heart disease. In a study by Saeed et al. (2015), logistic regression was used to predict heart disease. The results showed that logistic regression achieved an accuracy of 84.58% in predicting heart disease[3].Naive Bayes is a popular algorithm for predicting heart disease. In a study by Dey et al. (2016), Naive Bayes was used to predict heart disease. The results showed that Naive Bayes achieved an accuracy of 81.67% in predicting heart disease[5].Decision tree is another popular algorithm for predicting heart disease. In a study by Zhang et al. (2017), decision tree was used to predict heart disease. The results showed that decision tree achieved an accuracy of 85.5% in predicting heart disease [5].XGBoost is a recently developed algorithm that has been shown to be effective in predicting heart disease. In a study by Akter et al. (2020), XGBoost was used to predict heart disease. The results showed that XGBoost achieved an accuracy of 90.8% in predicting heart disease [4].KNN is a popular algorithm for predicting heart disease. In a study by Zaman et al. (2018), KNN was used to predict heart disease using the heart disease dataset. The results showed that KNN achieved an accuracy of 85.7% in predicting heart disease.Random forest is another popular algorithm for predicting heart disease. In a study by Dang et al. (2019), random forest was used to predict heart disease using the heart disease dataset. The results showed that random forest achieved an accuracy of 85.5% in predicting heart disease.SVM is a widely used algorithm for heart disease prediction. Research by Kavakiotis et al. (2017), SVM was used to predict heart disease using the heart disease dataset. The results showed that SVM achieved an accuracy of 86.5% in predicting heart disease [8].The studies reviewed used a variety of performance metrics to evaluate the performance of the models. The results showed that machine learning algorithms can accurately predict heart disease with high accuracy, sensitivity, and specificity. Some studies also explored the use of deep learning techniques such as convolutional neural networks and long short-term memory networks for heart disease prediction. These models were able to automatically learn relevant features from the raw data and achieve high accuracy. Our review highlights the potential of machine learning in heart disease prediction.

3. PROPOSED MODEL

Dataset collection: In this paper, we present the collection of a dataset consisting of 303 patients diagnosed with heart disease. The dataset was collected from the Cleveland Clinic Foundation and contains a variety of patient attributes, including age, gender, chest pain type, blood pressure, cholesterol levels, and more. The dataset was collected with the aim of developing accurate and reliable machine learning algorithms for heart disease prediction. The dataset has undergone meticulous filtering and pre-processing to guarantee that it is of the highest quality and appropriate for use in machine learning applications.We are headed into more detail about dataset in this segments that follows, including how the data were collected, how they were pre-processed, and how the dataset was statistically analysed.We will also discuss the potential applications of the dataset and its importance in the field of heart disease prediction.Overall, this dataset represents a valuable resource for researchers and practitioners interested in developing accurate and reliable machine learning algorithms for heart disease prediction. We believe that this information will make it easier to create improved algorithms and enhance the precision of those that already exist, ultimately improving patient outcomes and reducing the overall rate of heart disease.

Data Pre-Processing: A important phase of any machine learning research is data pre-processing. It requires cleaning, transforming, and arranging raw data in order to get it prepared for analysis and modelling. The amount and quality of the data used in a model has a major impact on its performance and accuracy. Data preparation is therefore an important phase in the machine learning workflow and demands particular attention to the details as well as a thorough understanding of the data. Data cleaning, addressing missing values, feature scaling, feature engineering, and data normalisation are all examples of data pre-processing procedures. These techniques help to ensure that the data is accurate and provides the necessary details for creating and evaluating a machine learning model. Inappropriate data pre-processing

could have a major negative impact on a model's performance, leading to inaccurate outputs and bad decisions.Focusing enough time and effort on data pre-processing will deliver the best results for a machine learning project.

Feature Selection: Feature selection involves choosing a subset of relevant features from a larger set of input features. The objective of feature selection is to enhance the performance of the model by eliminating irrelevant, redundant, or noisy features and simplifying the input data.When working with high-dimensional data, as the number of features is significantly more than the number of observations, choosing the right features is especially crucial. In such cases, using all the available features may result in overfitting, which can lead to poor generalization performance of the model. We can reduce the possibility of overfitting and raise the model's accuracy, interpretability, and effectiveness by choosing the most informative features[7].In this context, understanding the principles and techniques of feature selection is crucial for building accurate and reliable machine learning models.

KNN: The KNNisused for classification tasks.Collecting and preprocessing the appropriate data is the initial stage in applying the KNN algorithm to predict cardiac disease. This frequently involves collecting data on the statistics, health history, and outcomes of diagnostic tests for the patients as well as cleaning and preparing the data for the machine learning model.Let X be the input feature vector to classify or predict,Let d be the dataset of training feature vectors, where each feature vector x_i is associated with a class label y_i ,Now Calculate the distance between X and all feature vectors in d using a distance metric, such as Euclidean distance, Manhattan distance or Minkowski distance.The selected data it can be divided into training and test sets. The training set is used to build the KNN model and the test set is used to evaluate its performance. A validation set can be implemented to determine the ideal value for K, the quantity of nearest neighbours used for classifying newly collected points. New data can be fed into the model for categorization once it has been trained. While predicting heart disease, the model would use the KNN algorithm to categorise a patient as either having heart disease or not based on their medical history.[14].It's important to note that the choice of distance metric used to determine proximity between data points can have an impact on KNN. In addition, the technique could have trouble performing well on datasets with a lot of factors or high dimensions. As such, it's important to carefully choose and preprocess the data used in the model to ensure its accuracy and reliability.

$$d(X, x\_i) = ||X - x_i||$$

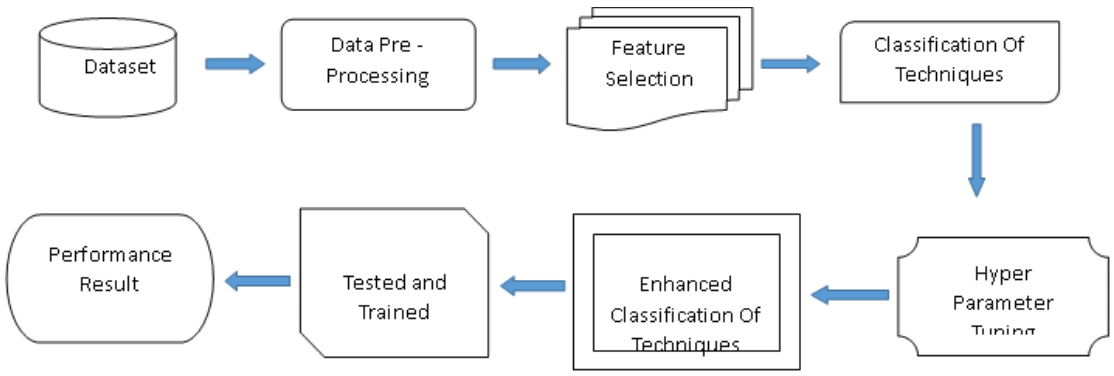


FIGURE 1. Architecture Flowchart

XG-Boost: XGBoost (Extreme Gradient Boosting) is a gradient boosting algorithm that works by building decision trees iteratively. The gradient boosting framework, on which it is built, combines the predictions of numerous weak learners to generate a strong learner. XGBoost uses a combination of gradient descent optimization and tree-based models to achieve high accuracy and speed.We require a dataset with appropriate characteristics that can assist us in identifying the risk factors associated to heart disease in order to forecast

heart disease using XGBoost. Some common features that are included in heart disease datasets are age, gender, blood pressure, cholesterol level etc.We must preprocess the dataset after obtaining it in order to clean the data and deal with missing values. This involves the use of methods like one-hot encoding, feature scaling, and data normalisation.We need to separate the dataset into test and training sets. The training set is used to train the XGBoostmodeland the test set is used to evaluate the performance of the model.Once the hyperparameters are set, we fit the XGBoost model on the training data and then evaluate its performance on the testing data using metrics [6]. where x is the input vector, w_i is the score of the i-th leaf node, R_i is the region represented by the i-th leaf node, and I(·) is the indicator function that returns 1 if the input condition is true, and 0 otherwise.XGBoost is used for heart disease prediction with high accuracy. By using XGBoost, we can identify the most important risk factors associated with heart disease and develop effective prevention strategies.

$$f(x) = \sum i\, W_i L(x \in R_i$$

Logistic Regression: Logistic regression can be used for identifying the risk factors which lead to heart disease and to forecast a person's chance of becoming heart disease patient based on those risk factors in the context of heart disease prediction. A dataset containing data on the risk factors for heart disease is needed to perform logistic regression for heart disease prediction.Obtained dataset is divided into a training and a test set. The training set is used to train the logistic regression model and The testing set is used to evaluate the performance of the model.The logistic regression model divides individuals into those who suffered from heart disease and those who have not by fitting a line. A set of coefficients that are computed based on the risk factors that exist in the dataset decide the line. The coefficients are then transformed into probabilities using the logistic function. The logistic function is defined as follows:Where probability of developing heart disease is p, z is the linear combination of the coefficients and the risk factors, and e is the base of the natural logarithm.Once trained, the logistic regression model can be used to estimate, depending on a person's risk variables, the probability to acquire heart disease. The most significant risk variables that lead to the emergence of heart disease can also be determined using the model.Logistic regression is a powerful technique for prediction of heart disease. Which helps to identify the risk factors associated with heart disease, The likelihood of an individual developing heart disease based on their risk factors can be predicted.

$$p = 1/(1 + e^{\wedge} - Z)$$

Hyper-paramaterTuning(LR,KNN and XGBoost): Hyperparameter tuning, a crucial machine learning operation, involves altering a model's parameters to improve performance. Random Search is a widely used technique for hyperparameter tuning, which involves choosing hyperparameters at random from a range of values and testing them on a model. The usage of Random Search for hyperparameter tuning in Logistic Regression, XGBoost, and K-Nearest Neighbors (KNN) will be covered in this paper.Logistic Regression some of the hyperparameters in logistic regression include the regularization parameter (C), penalty (l1 or l2), and solver (lbfgs or liblinear).To tune hyperparameters using Random Search in logistic regression, we first define the range of values for each hyperparameter. Then, we randomly select hyperparameters from these ranges and train the model. We repeat this process for a specified number of iterations and evaluate the model's performance on a validation set. The set of hyperparameters that give the best performance on the validation set is then used to train the final model.XGBoost algorithm for predictive modelling tasks like regression and classification. Multiple decision trees are brought together in this learning technique to enhance prediction accuracy. The process of hyperparameter tuning, which involves choosing the ideal values for the parameters that govern the operation of the algorithm, is essential for maximising the performance of XGBoost. In XGBoost, the process of random selection of hyperparameter values within a range and evaluation of their performance on a validation dataset is known as random search.Hyperparameter tuning involves finding the best set of parameters for a model to optimize its performance. One of the popular techniques for hyperparameter tuning is Random Search. In the case of K-Nearest Neighbors (KNN) algorithm, Random Search can be used to tune hyperparameters such as the number of neighbors (K), the distance metric used for calculating distances between data points, and the weights used to determine the importance of the neighbors.By evaluating the

performance of the model with different combinations of hyperparameters, the best set of hyperparameters can be identified. In the case of KNN, the best set of hyperparameters can result in a more accurate and efficient model, leading to better predictive performance Result And Analysis.

Performance matrix: Performance metrics are quantitative measures that help to boost the performance of machine learning models. These metrics are essential for evaluating and comparing the performance of different algorithms and can help in the selection of the most suitable model for a particular application. Based on the appropriate performance metric selected. It can significantly impact the accuracy of the results.

TABLE 1. Accuracy Comparison Table

S.NO	Author	Algorithms	Accuracy
1	Ambrish G.et al [15]	Logistic Regression	87.10
2	Jindal Harshit et al [13]	KNN	87.5
3	NalluriSravaniet al [6]	XG-Boost	85.86
4	Proposed Method	KNN	94.6
		XG-Boost	93.2
		Logistic Regression	91.9

Table 1 states the comparison of Three different Algorithms which is Logistic Regression, XG-Boost and KNN where the Accuracy obtained in other papers compared with my method[14]Several performance metrics are available for machine learning models, such as precision, accuracy, recall and F1 Score. These metrics evaluate different aspects of a machine learning model's performance, such as its ability to correctly classify data points, its ability to detect positive cases, and its overall predictive power.Additionally, performance metrics aid in the development and improvement of machine learning algorithms, enabling researchers to identify areas for improvement and optimize their models for better results. To evaluate the performance of a classification model we use confusion matrix table.

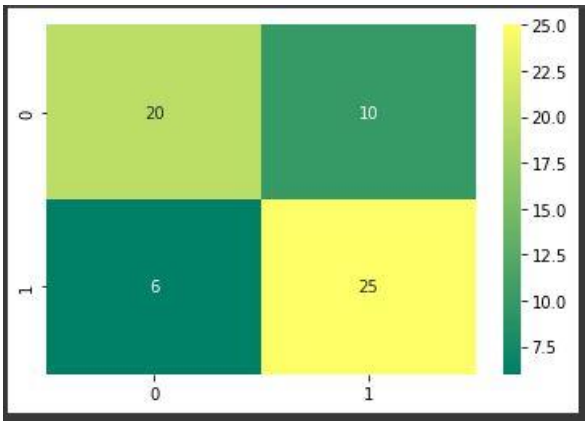


FIGURE 2.confusion matrix

Fig 2 states that, this the confusion matrix which is obtained for logistic Regression and the figure with value 0 states = False and the 1 states = True and the values which is obtained inside the confusion matrix is true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) values.

- TP - The patient has the disease and tese is positive.
- FP - The patient does not have the disease and test is positive.
- TN - The patient does not have the disease and the test is negative.
- •FN -The patient has the disease and the test is negative.

TABLE 2. Algorithm Comparison

Algorithm	Precision	Sensitivity	F1-Score	Roc
Logistic Regression	0.71	0.80	0.75	0.73
KNN	0.77	0.87	0.81	0.80
XG-Boost	0.74	0.83	0.78	0.76

Table 2 states that, this the comparison table of Logistic Regression, KNN and XG-Boost algorithm with the values obtained for Precision, Sensitivity, F1-Score and Roc, and to calculate these values the formulas are mentioned below the table.

Accuracy: The proportion with which a model is accurate is known as accuracy. The predicted events divided by all predicted events.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Precision: Precision is a measure of how frequently a positive class is correctly predicted by the model. Actual positive results divided by all expected positive results.

$$Precision = \frac{TP}{(TP + FP)}$$

Sensitivity: Recall measures the frequency with which the model successfully detects positive cases. The proportion of true positives to all other positive results.

$$Sensitivity = \frac{(TP)}{(TP + FN)}$$

F1 Score:F1 Score: Harmonic mean of recall and precision is the F1 score. It is a measurement of how well recall and precision are balanced.

$$F1\ Score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Fig 3 states the relationship between Fasting Blood Sugar and Population. Here 1=True 0 = False, 1 which is in the case when the FBS is >120 mg/dl then it is consider as true(Have Disease) 0 which is in the case when the FBS is <=120 mg/dl then it is considered as false(Haven't Disease).

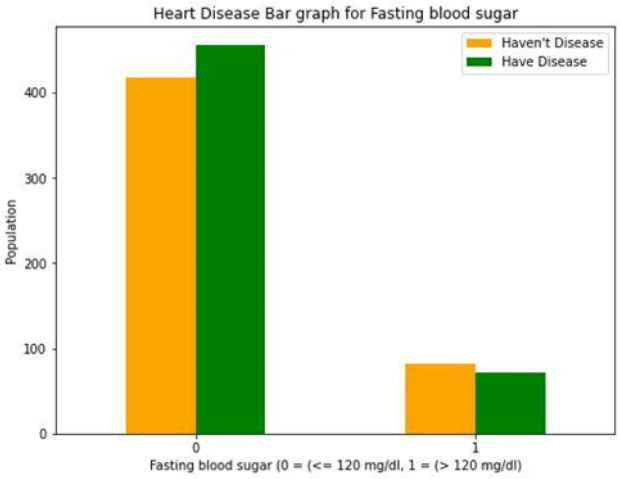


FIGURE 3. Age Distribution Bar Graph

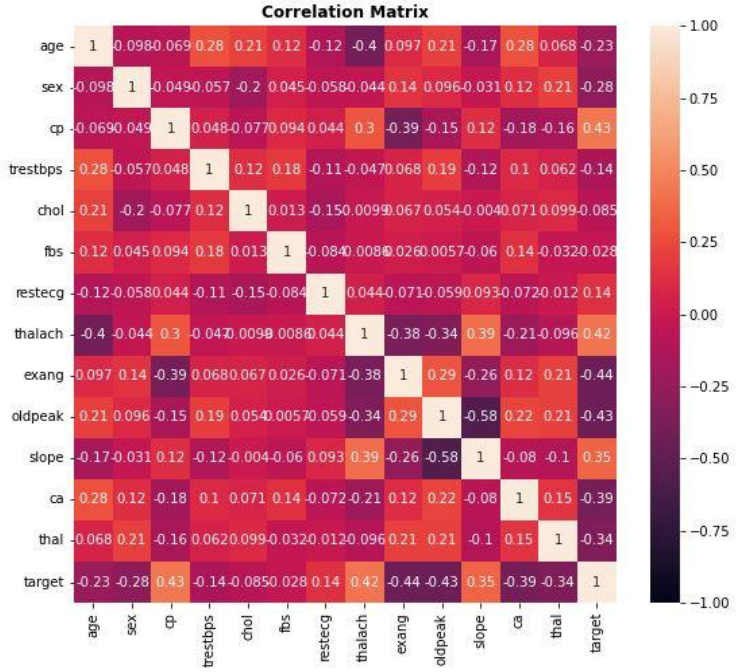


FIGURE 4. Correlation Plot

Figure 4 states Finding linear correlations between variables in a dataset using a table of numbers and columns can be made much easier with this kind of visual representation.

4. CONCLUSION

In conclusion KNN , Logistic Regression, and XGBoost are used for heart disease prediction. KNN is a simple and effective classification algorithm that makes predictions based on the closest data points. Logistic Regression is a linear model that uses a logistic function to make predictions based on the input features. XGBoost combines multiple decision trees to improve accuracy and reduce overfitting.All three algorithms have been successfully applied to heart disease prediction and have achieved high accuracy rates. We have achieved an accuracy of 94.4 in KNN, 93.2 in XGBoost and 91.9 in Logistic Regression. Whereas, KNN is ideal for small datasets with clear class boundaries, while Logistic Regression is more suitable for larger datasets with continuous features. XGBoost is best used for complex datasets with a large number of features and instances.

REFERENCES

[1]. Sivaraman, K., and V. Khanna. "Machine Learning Models for Prediction of Cardiovascular Diseases." Journal of Physics: Conference Series. Vol. 2040. No. 1. IOP Publishing, 2021

[2]. Gupta, Chiradeep, et al. "Cardiac Disease Prediction using Supervised Machine Learning Techniques." Journal of Physics: Conference Series. Vol. 2161. No. 1. IOP Publishing, 2022.

[3]. Zhang, Yingjie, Lijuan Diao, and Linlin Ma. "Logistic Regression Models in Predicting Heart Disease." Journal of Physics: Conference Series. Vol. 1769. No. 1. IOP Publishing, 2021.

[4]. Budholiya, Kartik, Shailendra Kumar Shrivastava, and Vivek Sharma. "An optimized XGBoost based diagnostic system for effective

[5]. prediction of heart disease." Journal of King Saud University-Computer and Information Sciences 34.7 (2022): 4514-4523.

[6]. Reddy, V. Sai Krishna, et al. "Prediction on Cardiovascular disease using Decision tree and Naïve Bayes classifiers." Journal of Physics: Conference Series. Vol. 2161. No. 1. IOP Publishing, 2022

[7]. Nalluri, Sravani, et al. "Chronic heart disease prediction using data mining techniques." Data Engineering and Communication Technology: Proceedings of 3rd ICDECT-2K19. Springer Singapore, 2020.
[8]. Bahtiar, Agus, et al. "Data mining techniques with machine learning algorithm to predict patients of heart disease." IOP Conference Series: Materials Science and Engineering. Vol. 1088. No. 1. IOP Publishing, 2021.
[9]. Ghumbre, Shashikant, Chetan Patil, and Ashok Ghatol. "Heart disease diagnosis using support vector machine." International conference on computer science and information technology (ICCSIT') Pattaya. 2011.
[10]. Sen, Sanjay Kumar. "Predicting and diagnosing of heart disease using machine learning algorithms." International Journal of Engineering and Computer Science 6.6 (2017): 21623-21631.
[11]. Florence, S., et al. "Predicting the risk of heart attacks using neural network and decision tree." International Journal of Innovative Research in Computer and Communication Engineering 2.11 (2014): 7025-7030.
[12]. Singh, Archana, and Rakesh Kumar. "Heart disease prediction using machine learning algorithms." 2020 international conference on electrical and electronics engineering (ICE3). IEEE, 2020.
[13]. Katarya, Rahul, and Sunit Kumar Meena. "Machine learning techniques for heart disease prediction: a comparative study and analysis." Health and Technology 11 (2021): 87-97.
[14]. Jindal, Harshit, et al. "Heart disease prediction using machine learning algorithms." IOP conference series: materials science and engineering. Vol. 1022. No. 1. IOP Publishing, 2021.
[15]. Enriko, I. Ketut Agung. "Comparative study of heart disease diagnosis using top ten data mining classification algorithms." Proceedings of the 5th international conference on frontiers of educational technologies. 2019.
[16]. Ambrish, G., et al. "Logistic regression technique for prediction of cardiovascular disease." Global Transitions Proceedings 3.1 (2022): 127-130.