Abstract

Heart disease diagnosis often requires a complex combination of clinical and pathological data, leading to increased healthcare costs and reduced quality of care. With heart disease being a leading cause of mortality globally, efficient and accurate prediction systems are crucial for early detection and intervention. This study explores the application of machine learning algorithms, including logistic regression, decision trees, XGBoost, and support vector machines (SVM), in predicting heart disease based on clinical and pathological data. Using a dataset spanning multiple databases and containing 14 attributes, including key predictors such as age, sex, chest pain type, resting blood pressure, serum cholesterol, and other relevant factors, predictive models were trained and evaluated. Logistic regression emerged as the most effective algorithm, achieving a test accuracy of 90.16%. Decision tree, XGBoost, and SVM followed closely with accuracies of 86.89%, 83.61%, and 85.25%, respectively. Logistic regression's superior performance makes it a promising tool for heart disease prediction in clinical settings, enabling healthcare professionals to make informed decisions about diagnosis, treatment, and patient management. By leveraging machine learning techniques, such as logistic regression, healthcare providers can enhance patient care, improve outcomes, and mitigate the burden of heart disease on public health.

Keywords: Heart disease, machine learning algorithms, clinical data, predictive modelling, patient management, healthcare professionals.

1.0 Introduction

In most cases, a complex combination of clinical and pathological data is required to diagnose heart disease; this complexity raises the cost of care excessively and lowers its quality (Wu, Peter and Morgan, 2002). According to WHO statistics, heart disease claimed the lives of one-third of the global population in 2010 and was the leading cause of death in developing nations.

Chen et al. (2011) state that the prediction of heart disease and HDPS indicates the application of a number of methods and algorithms, such as data mining, decision trees, neural networks, and Naive Bayes, to create prediction systems. Using clinical and pathological data, these systems are designed to help with the efficient and accurate prediction of heart disease.

Large data sets can be analysed by machine learning algorithms, which can also spot patterns that humans might find difficult to notice. Machine learning models can accurately predict the presence or risk of heart diseases by utilising these patterns (Bharti and associates, 2021). For prompt intervention and treatment of heart diseases, early detection is essential. To determine who is more likely to develop heart disease, ML models can analyse a variety of risk factors and clinical data. This makes it possible for medical professionals to take preventative action and intervene early.

Bhatt et al. (2023) state that these models can support clinical decision-making by offering predictions and insights based on patient data, enabling medical professionals to make more educated decisions about diagnosis, treatment, and patient management. These models can assist researchers in developing new prevention and treatment strategies as well as a deeper understanding of the disease by analysing large datasets and finding new patterns and risk factors.

2.0 Literature Review

The table below show the data set clinical features and their description.

2.1 Data Source

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 14 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.

2.2 Features and Description

Table 1: The table below shows the features and description of the dataset

Features	Description
Age	Age in years
Sex	1=male, 0=female
cp	Chest pain type 0=typical, 1=atypical,2= non-angina,3= asymptomatic
trestbps	Resting blood pressure in mmHg
chol	Serum cholesterol in mg/dl
fbs	Fasting blood sugar>120 mg/dl 1=true, 0=false
restecg	Resting electrocardiographic result
thalach	Maximum heart rate achieved
Exang	Exercise induced angina 1=yes, 0=no
Oldpeak	ST depression induced by exercise relative to rest
slope	The slope of the peak exercise ST segment
ca	Number of major vessel (0-3) coloured
thal	thal: 0 = normal ,1 = fixed defect, 2 = reversible defect
target	0= without heart disease 1= with heart disease

2.3 Logistic Regression

When the explanatory variables are continuous but the response variable is binary, logistic regression can be helpful. This would be the case if data on an individual's income, years of employment, age, education, and other continuous variables were used to predict whether or not the customer is a good credit risk.

$$PY = 1 = \frac{expX^{T}\theta}{1+expX^{T}\theta} \quad \textit{Equation 1}$$

Where Y is equal to If a customer is deemed a good risk, θ represents the unknown parameters that need to be estimated from the data, and X is the vector of explanatory variables for that customer. The benefit of this model is that, under the transformation.

$$p = ln\frac{P[Y=1]}{1-p[Y=1]} \textit{Equation 2}$$

The linear model p = $X^{T}\theta$ is obtained. As a result, all of the standard multiple linear regression tools will work.

By creating dummy variables, logistic regression can be adjusted to handle categorical explanatory variables; however, this becomes unfeasible when there are numerous categories.

2.4 Decision tree

The Decision Tree algorithm forecasts a problem's result by using a data structure known as a tree. A set of pre-processed data is fed into the algorithm because the decision tree uses a supervised methodology. The algorithm is trained with this data. Find out more information about this by going here. The main concept is to divide the data space into dense and sparse regions using a decision tree. Binary or multiday splitting of a binary tree is possible. The tree is kept splitting by the algorithm until the data is fairly homogeneous. Upon completion of training, an optimal categorised prediction can be made using the decision tree that is returned.

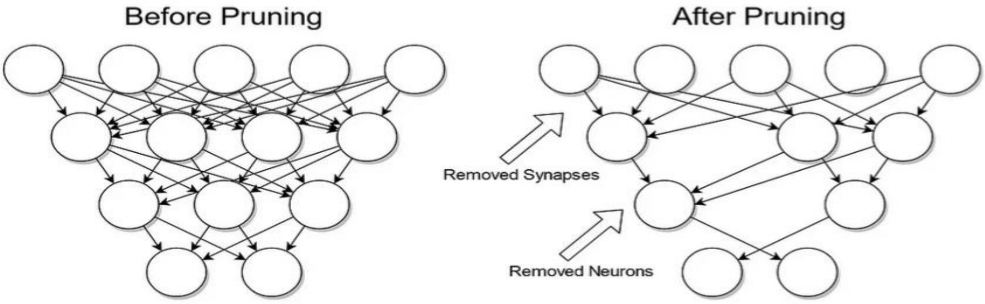


Figure 1: Decision Tree Pruning and Post-Pruning

This kind of scenario arises when there is very little margin of error for a result, which makes the model doubt the prediction's accuracy. The randomness in the dataset will increase with increasing entropy. It is desirable to have a lower entropy when creating a decision tree. The following is the formula for determining a decision tree's entropy:

$$Entropy = \sum_{i=1}^{c} - p_i * log_2(p_i) \quad Equation\ 3$$

2.5 Xgboost

An ensemble learning machine learning algorithm is called eXtreme Gradient Boosting, or XGBoost. For supervised learning tasks like regression and classification, it's in. By iteratively combining the predictions of several different models, frequently decision trees, XGBoost creates a predictive model.

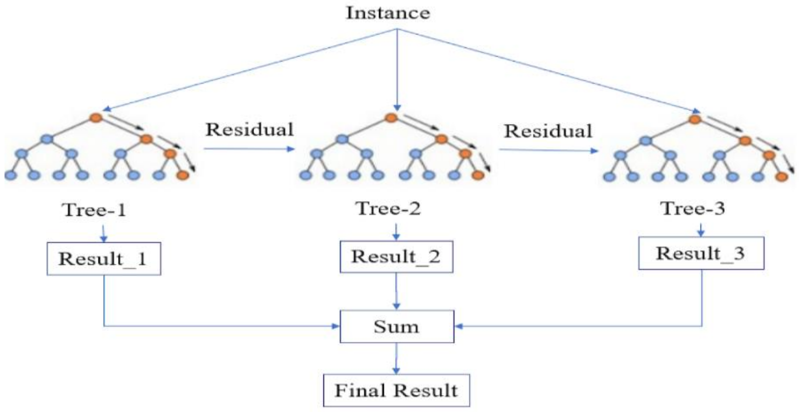


Figure 2: XGBoost Algorithm Overview

In order for the algorithm to function, weak learners are gradually added to the ensemble, with each new learner concentrating on fixing the mistakes made by the previous ones. Throughout training, it minimises a predetermined loss function by using an optimisation technique called gradient descent.

2.6 Support vector machine

The goal of the support vector machine algorithm is to find a hyper plane in an N-dimensional space (where N is the number of features) that categories the data points clearly. There are numerous hyper planes that could be selected in order to divide the two classes of data points apart. Finding a plane with the maximum margin—that is, the maximum separation between data points for both classes—is our goal. In order to classify subsequent data points more confidently, it is helpful to maximise the margin distance.

3.0 Methodology

Our approach consists of multiple interrelated phases with the goal of creating reliable and accurate heart disease prediction models.

The first step entails obtaining a representative and diverse dataset that includes clinical measurements pertinent to heart health, medical history, and demographic data. To ensure data quality and compatibility with machine learning algorithms, the data are then pre-processed to fix missing values, dropping duplicates, encode categorical variables, and normalise numerical features.

To assess the performance of predictive models, the dataset is split into training and testing sets. 80% of the dataset is used to train the model while 20% of the dataset is used to evaluate their performance. Following data splitting, the training set is used to train predictive models using a selection of algorithms, including Logistic Regression, Decision Tree, XGBoost, and Support Vector Machine (SVM).

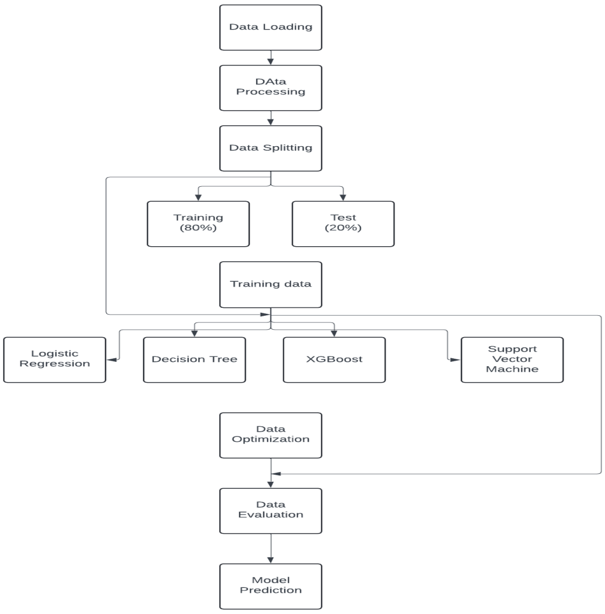


Figure 3: The steps of the heart diseases model prediction

After training, the models undergo optimization to fine-tune their parameters and improve performance. Techniques such as grid search was used to explore the hyper parameter space and identify optimal configurations for each algorithm.

Evaluation metrics such as accuracy score, recall and F1-score are employed to assess the performance of trained models on the testing set. Once trained and evaluated, the predictive models are ready for deployment and use in making predictions on new, unseen data. By inputting relevant features into the trained models, predictions can be generated, providing valuable insights and forecasts for decision-making purposes.

4.0 Result

4.1 Data Visualization

The range of the resting blood pressure (trestbps) falls between 80 and 220, with the mode typically between 120 and 140. The range of serum cholesterol (chol) falls between 100 and 600, with the mode usually between 200 and 300. The features cp, restecg, slope, and thalach are positively correlated with the target, while the remaining features are negatively correlated. There are a total of 206 male patients, with up to 90 of them having heart diseases, while the total number of female patients is 96, with up to 70 of them having heart diseases. 45 patients have fasting blood sugar, while 257 do not. 143 patients have typical chest pain, 86 have atypical chest pain, 50 have non-angina chest pain, and 23 have asymptomatic chest pain.

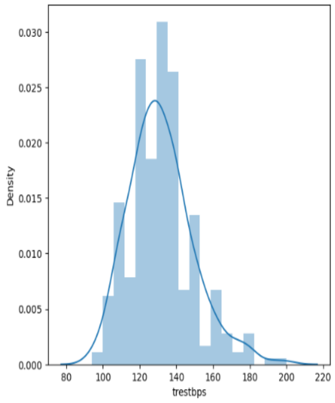


Figure4: trestbps range

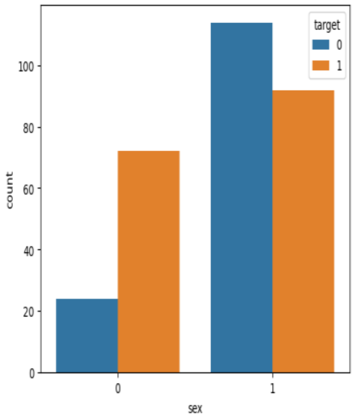


Figure5: Sex value count

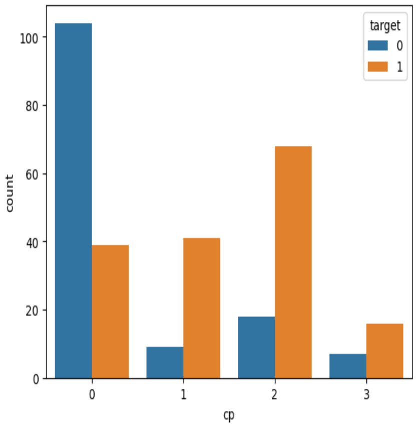


Figure6: cp value count

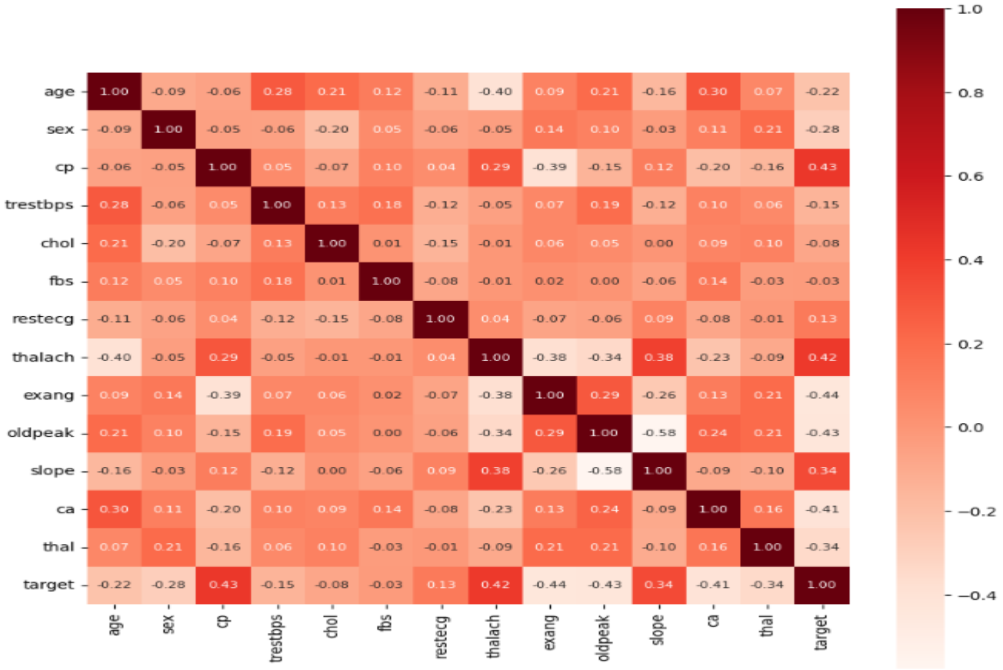
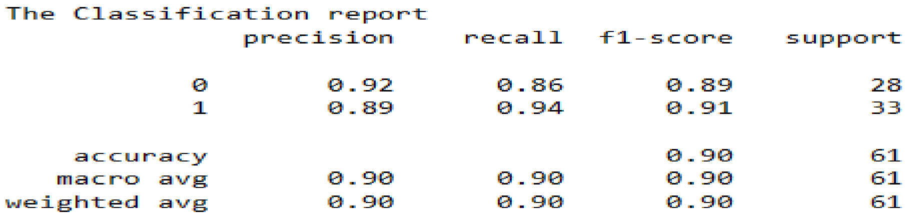


Figure 7: The correlation map of the heart disease dataset

4.2 Logistic Regression

The classification report indicates an overall accuracy of 90% for the model's predictions shown in figure 8. It achieved a high precision of 92% for class 0 and 89% for class 1, indicating a low false positive rate. The recall, or sensitivity, is 86% for class 0 and 94% for class 1, showing the model's ability to capture true positives. The F1-score, which balances precision and recall, is 89% for class 0 and 91% for class 1, demonstrating the model's effectiveness in both classes.

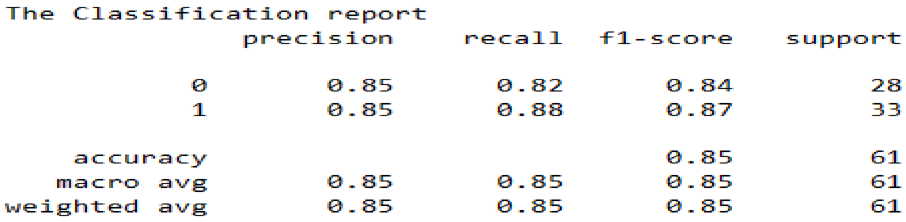


```
The Classification report
              precision    recall  f1-score   support

           0       0.92      0.86      0.89        28
           1       0.89      0.94      0.91        33

    accuracy                           0.90        61
   macro avg       0.90      0.90      0.90        61
weighted avg       0.90      0.90      0.90        61
```

Figure 8: The classification report of the logistic regression model

4.3 Decision Tree

The classification report presents an accuracy of 85% for the model's predictions, indicating its overall performance show in figure 9. Precision values of 85% for class 0 and 85% for class 1 suggest a balanced ability to correctly identify instances of each class. The recall values of 82% for class 0 and 88% for class 1 illustrate the model's capability to capture true positives effectively. F1-scores of 84% for class 0 and 87% for class 1 demonstrate a harmonized balance between precision and recall for both classes.

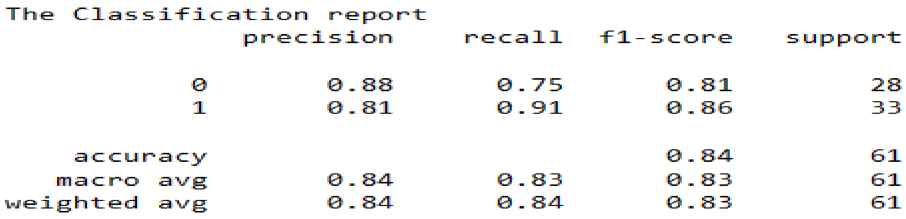


```
The Classification report
              precision    recall  f1-score   support

           0       0.85      0.82      0.84        28
           1       0.85      0.88      0.87        33

    accuracy                           0.85        61
   macro avg       0.85      0.85      0.85        61
weighted avg       0.85      0.85      0.85        61
```

Figure 9: The classification report of the decision tree model

4.3 XGBoost

The classification report illustrates an accuracy of 84% for the model's predictions, indicating its overall performance show in figure 10. Precision values of 88% for class 0 and 81% for class 1 demonstrate the model's ability to correctly identify instances of each class, albeit with some variance. Recall values of 75% for class 0 and 91% for class 1 suggest a stronger capability to capture true positives for class 1. F1-scores of 81% for class 0 and 86% for class 1 indicate a balanced performance between precision and recall for both classes, though slightly favouring class 1.

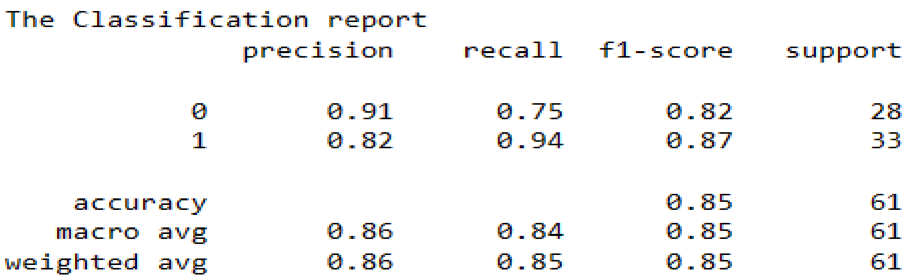


```
The Classification report
              precision    recall  f1-score   support

           0       0.88      0.75      0.81        28
           1       0.81      0.91      0.86        33

    accuracy                           0.84        61
   macro avg       0.84      0.83      0.83        61
weighted avg       0.84      0.84      0.83        61
```

Figure 10: The classification report of the XGBoost model

4.4 Support Vector Machine (SVM)

The classification report demonstrates an accuracy of 85% for the model's predictions, indicating strong overall performance show in figure 11. Precision values of 91% for class 0 and 82% for class 1 highlight the model's ability to accurately identify instances of each class. Recall values of 75% for class 0 and 94% for class 1 indicate the model's effectiveness in capturing true positives, particularly for class 1. F1-scores of 82% for class 0 and 87% for class 1 suggest a harmonized balance between precision and recall for both classes.



```
The Classification report
              precision    recall  f1-score   support

           0       0.91      0.75      0.82        28
           1       0.82      0.94      0.87        33

    accuracy                           0.85        61
   macro avg       0.86      0.84      0.85        61
weighted avg       0.86      0.85      0.85        61
```

Figure 11: The classification report of the SVM model

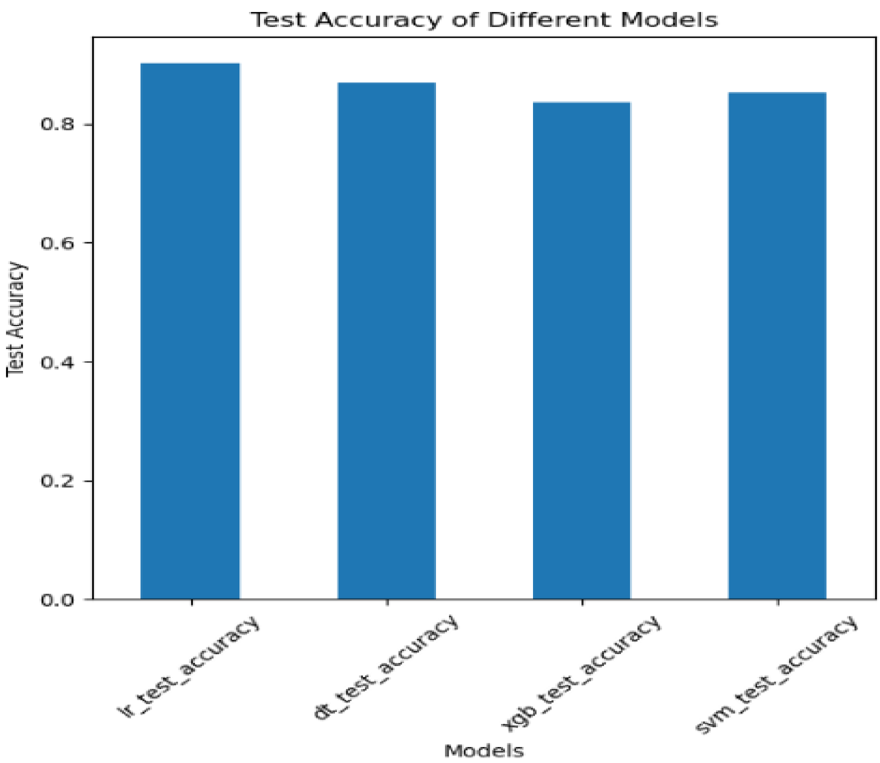


Figure 10: The test accuracy of different models

5.0 Conclusion

Based on the accuracy results obtained from various machine learning algorithms as shown in figure 11, it's evident that logistic regression outperforms the other models with a test accuracy of 90.16%. Following logistic regression, decision tree achieved a test accuracy of 86.89%, while xgboost and support vector machine attained accuracies of 83.61% and 85.25% respectively.

Considering these findings, logistic regression emerges as the most suitable algorithm for predicting heart disease in this dataset. Its higher accuracy indicates better performance in classifying instances correctly compared to the other models.

For future use, logistic regression can serve as a reliable tool for heart disease prediction in clinical settings. Its high accuracy suggests it can assist healthcare professionals in making informed decisions about patient diagnoses and treatment plans.

Reference

1. Wu, R., Peters, W. and Morgan, M.W., 2002. The next generation of clinical decision support: linking evidence to best practice. Journal of healthcare information management: JHIM, 16(4), pp.50-55.
2. Chen, A.H., Huang, S.Y., Hong, P.S., Cheng, C.H. and Lin, E.J., 2011, September. HDPS: Heart disease prediction system. In *2011 computing in Cardiology* (pp. 557-560). IEEE
3. Bhatt, C.M., Patel, P., Ghetia, T. and Mazzeo, P.L., 2023. Effective heart disease prediction using machine learning techniques. Algorithms, 16(2), p.88.
4. Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S. and Singh, P., 2021. Prediction of heart disease using a combination of machine learning and deep learning. Computational intelligence and neuroscience, 2021.
5. Palaniappan, S. and Awang, R., 2008, March. Intelligent heart disease prediction system using data mining techniques. In 2008 IEEE/ACS international conference on computer systems and applications (pp. 108-115). IEEE.
6. Saxena, K. and Sharma, R., 2016. Efficient heart disease prediction system. Procedia Computer Science, 85, pp.962-969.
7. Dangare, C.S. and Apte, S.S., 2012. Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), pp.44-48.
8. Parthiban, L. and Subramanian, R., 2008. Intelligent heart disease prediction system using CANFIS and genetic algorithm. International Journal of Biological, Biomedical and Medical Sciences, 3(3).
9. Taneja, A., 2013. Heart disease prediction system using data mining techniques. Oriental Journal of Computer science and technology, 6(4), pp.457-466.
10. Subbalakshmi, G., Ramesh, K. and Rao, M.C., 2011. Decision support in heart disease prediction system using naive bayes. Indian Journal of Computer Science and Engineering (IJCSE), 2(2), pp.170-176.
11. Rajamhoana, S.P., Devi, C.A., Umamaheswari, K., Kiruba, R., Karunya, K. and Deepika, R., 2018, July. Analysis of neural networks based heart disease prediction system. In 2018 11th international conference on human system interaction (HSI) (pp. 233-239). IEEE.