

HEART DISEASE IDENTIFICATION AND NOTIFICATION SYSTEM USING MACHINE LEARNING

Aishwarya B, ²Deepika A, ³Harithaa S, ⁴Narmadha D, ⁵Ms. Lavanya. S

¹Student, ²Student, ³Student, ⁴Student, ⁵Assistant Professor

¹Department of Computer Science and Engineering,

¹KCG College of Technology, Chennai, India

Abstract— The most serious health problem is heart disease (HD), which has affected many people all over the world. Shortness of breath, muscular weakness, and swollen feet are prominent signs of HD. Due to a number of factors, including accuracy and execution time, present heart disease diagnosis techniques are not very effective in early time identification. Researchers are working to develop an effective method for the detection of heart disease. In this study, we suggested a machine learning-based system that can quickly and accurately diagnose cardiac problems. The categorization techniques used in the system's development include support vector machines, decision trees, and random forests. Then, in order to increase accuracy, we combined all of the algorithms mentioned earlier to create a hybrid algorithm. The classifiers' performances are evaluated using the performance measurement metrics. On the features chosen via features selection algorithms, the classifiers' performances have been evaluated. The flask framework website provides the final forecast. The suggested system can also recommend food to patients who test positive.

keywords-- support vector machines, decision trees, random forests, hybrid algorithm, performance measurement metrics, feature selection, flask framework, food recommendations.

I. INTRODUCTION

The World Health Organization estimates that heart disease causes 12 million deaths worldwide each year. One of the leading causes of morbidity and mortality among the global population is heart disease. One of the most crucial topics in the data analysis area is predicted cardiovascular disease. Since a few years ago, the prevalence of cardiovascular disease has been rising quickly throughout the world. Numerous studies have been undertaken in an effort to identify the causes of heart disease that have the greatest impact and to precisely forecast the overall risk. Because it results in death without any overt symptoms, heart disease is also known as the "silent killer." Early detection of heart disease in high-risk persons is essential for assisting them in deciding whether to alter their lifestyle, which reduces consequences.

Making choices and predictions from the vast amounts of data generated by the healthcare sector is made easier with the help of machine learning. This study aims to forecast future cases of heart disease by analyzing patient data that makes use of a machine-learning system to categorize whether a patient has heart disease or not. In this case, machine learning techniques can be of great assistance. Despite the fact that heart disease can present itself in a variety of ways, there is a common set of fundamental risk factors that determine whether or not someone would ultimately be at risk for the condition. By compiling data from many sources, organizing them under the proper headings, and then analyzing the data to extract the necessary information, we can say that this technique can be very effectively used to forecast heart disease.

Machine learning is important because it enables the creation of new products and gives organizations an understanding of consumer behavior trends and operative business patterns. Machine learning is crucial to the operations of many of the leading companies of today, like Facebook, Google, and Uber. Machine learning has become a major point of competitive difference for many firms. In supervised learning, data scientists describe the variables they want the algorithm to look for connections between and provide the algorithms with labelled training data. The algorithm's input and output are both described. Algorithms used in unsupervised learning are trained on unlabeled data. The algorithm looks for any meaningful relationships among data sets. Both the predictions or suggestions that algorithms generate as well as the data that they use to learn are preset. A machine learning technique that combines the two aforementioned categories is called semi-supervised learning.

Data scientists may provide an algorithm with mostly labelled training data, but the algorithm is still free to explore the data on its own and draw its own conclusions about the data set. Reinforcement learning is widely used by data scientists to teach a computer to carry out a multi-step procedure for which there are predefined rules. Data scientists design an algorithm to achieve a goal, and when it chooses how to do so, they provide it good or negative feedback. However, the algorithm usually makes the decision on its own. The data scientist must use labelled inputs and desired outputs to train the algorithm in supervised machine learning. For the tasks listed below, supervised learning techniques work well: By categorizing data into two groups, binary classification. Multiple-class classification: Choosing from among more than two sorts of responses. Making predictions for continuous values using regression modelling. Assembling: Compiling the results of various machine learning models to obtain a precise prediction.

I. RELATED WORKS

Numerous studies and tests have been conducted in recent years, with the relevant major articles being published, with expanding advancements in the fields of machine learning and medical science.

Research titled "Efficient Heart Disease Prediction System" by Purushottam et al. [1] advocated employing algorithms that use decision trees and hill climbing. They used the Cleveland dataset, and they pre-processed the data before applying classification techniques. The Knowledge Extraction method is built on the open-source Evolutionary Learning (KEEL) data mining technique, which fills in the missing values in the data set. A top-down approach is used when using a decision tree. For each actual node selected by the hill-climbing algorithm at each level, a node is selected by a test. The variables and their corresponding values are confidence. Its level of confidence is at least 0.25. The accuracy of the system is about 86.7% of the time.

In their study "Prediction of Heart Disease Using Machine Learning Algorithms," Santhana Krishnan, J., et al. [2] suggested using decision trees and the Naive Bayes method to predict heart disease. The True or False options generated by the decision tree algorithm are used to form the tree. Based on split criteria, which can be vertical or horizontal depending on the dependent variables, algorithms like SVM and KNN produce their results. However, a decision tree is a structure that resembles a tree and is based on the choices made in each tree. It has a root node, leaves, and branches. The value of the attributes in the dataset is also explained by the decision tree. Additionally, they used the Cleveland data set. Using some techniques, the data set is divided into 70% training and 30% testing. The accuracy of this method is 91%. Naive Bayes, the second algorithm, is used for categorization. Since it can handle complex, nonlinear, dependent data, the heart disease dataset—which is similarly complex, dependent, and nonlinear in nature—is seen to be a good fit. 87% accuracy is provided by this method.

In their study "Prediction of Heart Disease Using Machine Learning Algorithms," Sonam Nikhar et al. [3] propose that Naive Bayes and decision tree classifier, which are utilized specifically in the prediction of Heart Disease, are explained in detail. Studies that looked at using a predictive data mining strategy on the same dataset found that Decision Trees had a higher accuracy than Bayesian classifiers.

In their publication "Prediction of Heart Disease Using Machine Learning," Aditi Gavhane et al. [4] advocated utilizing the multi-layer perceptron neural network approach to train and test datasets. There will be one input layer, one output layer, and perhaps more hidden layers in this algorithm between the two input and output layers. Each input node is connected to the output layer by hidden layers. Weights chosen at random are assigned to this link. The second input is referred to as bias, and it is given weight according to the needs of the connection between the nodes.

In their proposal "Heart Disease Prediction Using Effective Machine Learning Techniques," Avinash Golande et al. [5] used a variety of data mining techniques, including packing calculation, part thickness, consecutive negligible streamlining, neural systems, straight Kernel self-argument, and Nave Bayes, to help doctors distinguish between different types of heart disease. There may be more causes of heart disease, according to Lakshmana Rao et al. in their article "Machine Learning Techniques for Heart Disease Prediction" [6]. It is so difficult to discern between different types of cardiac disease. The severity of cardiac disease among patients is determined using a variety of neural networks and data mining techniques.

The "Heart Attack Prediction Using Deep Learning" proposal by Abhay Kishore et al. [7] calls for the use of a Recurrent Neural System to forecast the likelihood of heart-related infections in patients and the usage of a Deep Learning-based heart attack prediction system. For the best accuracy and fewest errors, this model combines deep learning and data mining. For many heart attack prediction models, this study serves as a reliable reference model.

Senthil Kumar Mohan et al. [8] presented "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" to improve cardiovascular issue accuracy. The increased demonstration level with an accuracy level of 88.7% is produced by the prediction model for heart disease leveraging hybrid random forest with linear model and using the algorithms KNN, LR, SVM, and NN.(HRFLM).

A model that explained and contrasted the performance of prediction for two categorization models was developed by Anjan N. Repaka and colleagues [9]. The experimental results show that our proposed strategy outperforms other models in terms of accuracy in forecasting the percentage of risk.

According to Aakash Chauhan et al. in "Heart Disease Prediction using Evolutionary Rule Learning," electronic records enable direct data retrieval, minimizing the requirement for manual processes. To aid with the best heart disease prognosis, the number of services is decreased and a big number of regulations are displayed. Strong associations are produced using frequent pattern growth association mining on the patient's dataset.

II. METHODOLOGY

The suggested approach makes an effort to predict a patient's chance of having heart disease based on a number of factors. To display the distribution of the variables, density maps are first plotted using Matplotlib. The correlation between the variables is then displayed using a heatmap from the seaborn library. The preparation of the data for the model-building process is the following step. The categorical variables were converted into the binary variables "cp," "thal," "slope," and "gender" using one-hot encoding. The old category variables are eliminated, and the old data is mixed with the old data. The data is then scaled using the StandardScaler function from the sklearn library. Next, training and testing sets are created from the data using the train_test_split

function from the sklearn library. Training and testing data are used to develop three different machine learning algorithms: support vector machine, decision tree, and random forest. Then, using a hybrid method that combined all of these algorithms, we improved accuracy. The matrices for accuracy, precision, recall, and confusion are printed for each model. The model's performance is assessed using the accuracy, recall, and confusion matrices, which are also used to comprehend the true positive, false positive, true negative, and false negative rates. The last web prediction/food recommendation for people with heart disease.

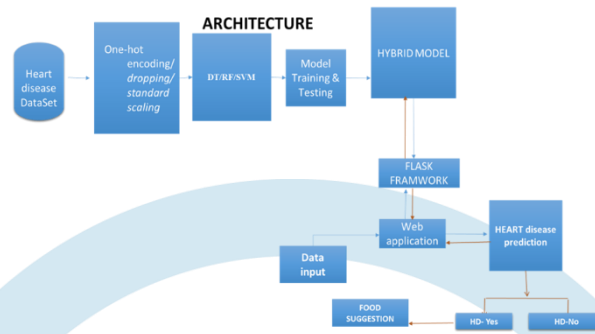


Fig 1: Architectural diagram of the proposed system

A. DATA COLLECTION:

Building a machine learning-based system for detecting heart disease requires the collecting of data. A sizable dataset of patient data is necessary, including information from imaging investigations, physical exams, blood test results, and medical histories. Data must be correct, dependable, and population-representative. The right steps must be taken, like anonymizing patient data and adhering to legislation, to protect data security and privacy. To enhance the quality of the data and get it ready for machine learning algorithms, data pre-processing techniques including cleaning, normalization, and feature extraction can be used. A cardiac disease identification system must collect data carefully and completely to be successful.

S.no	DATA	DESCRIPTION	0	1	2	3	4	5	6	7
1	age	Displays the age of interval								
2	sex	displays the gender of the individual using the following format	female	Male						
3	chest pain type	displays the type of chest-pain experienced by the individual using the following format		Typical angina	Atypical angina	Non-anginal pain	Asymptotic			
4	resting blood pressure	displays the resting blood pressure value of an individual in mmHg (unit)								
5	serum cholesterol	displays the serum cholesterol in mg/dl (unit)								
6	fasting blood sugar	compares the fasting blood sugar value of an individual with 120mg/dl.	false	Blood sugar>120mg/dl: true						
7	resting ecg	displays resting electrocardiographic results	Normal	Having ST-T wave abnormality	Left ventricular hypertrophy					
8	max heart rate achieved	displays the max heart rate achieved by an individual.								
9	exercise induced angina		no	yes						
10	st depression induced by exercise relative to rest	displays the value which is an integer or float.								
11	peak exercise st segment			upsloping	flat	down sloping				
12	number of major vessels (0-3) colored by fluoroscopy	displays the value as integer or float.								
13	thal	displays the thalassemia				Normal			Fixed defect	Reversible defect
14	diagnosis of heart disease	Displays whether the individual is suffering from heart disease or not	absent	present	present	present	present			

Fig 2 : Data Collection table

B. DATA PRE-PROCESSING:

Categorical Variables Into Binary Variables Using One-Hot Encoding, StandardScaler()Test Train Split ,Data Analysis, Drop Id.

C. MODEL TRAINING:

This entails feeding the model input data (also referred to as features) and the appropriate output data (often referred to as labels), then letting the model learn from the data by modifying its parameters. There are several machine learning algorithms from which to pick. The algorithm chosen will depend on the data's makeup and the problem being solved. A potent machine learning technique called Random Forest (RF) has been successfully used to forecast cardiac disease. In this situation, RF is utilised to model the association between a set of predictor factors (such as age, gender, blood pressure, cholesterol levels, and family history of heart disease) and the binary outcome of whether or not a patient has heart disease.

RF is a particular kind of ensemble learning technique that constructs numerous decision trees and combines their output to produce predictions. To assist prevent overfitting and improve the model's generalizability, each tree is trained using a random portion of the data and a random subset of the predictor variables.

The use of RF in the diagnosis of cardiac disease has various benefits:High accuracy: When compared to other modelling methods like logistic regression, RF models have been found to have a high accuracy in predicting heart disease.

RF can manage complicated interactions between predictor variables, which may be challenging to model using conventional statistical methods.RF is resistant to noise and missing data, which are common issues in datasets used in the medical field.Can handle categorical and continuous predictor variables: RF is a versatile modelling technique that can handle categorical and continuous predictor variables.

SVM :Due to its many benefits, Support Vector Machines (SVM), a popular machine learning technique, can be used to predict cardiac disease. SVM has been demonstrated to be more accurate than other machine learning algorithms at predicting heart disease. Because non-linear correlations in the data, which are frequently present in medical datasets, are well-handled by SVM, this is the case. Overfitting, a typical issue with machine learning algorithms, is a problem that SVM is comparatively resistant to. SVM is a good method for predicting heart disease since it can handle high dimensional data and generalise well to fresh data. SVM establishes a distinct line between classes, making it simple to comprehend how the algorithm generates its predictions.

DECISION TREE:Another well-liked machine learning approach that can be used to predict heart disease is the decision tree (DT).

Scalability: DT is suited for large medical datasets since it can handle datasets with thousands of observations and hundreds of features.

Interactivity: DT can be utilised in a way that allows the user to adjust the decision criteria in order to explore various scenarios and comprehend how various features affect the prediction.DT can be integrated with other algorithms employing ensemble techniques, such as Random Forests, to increase prediction accuracy.

HYBRID ALGORITHM: A set of input features (X) and three models—a Decision Tree (DT) model, a Random Forest (RF) model, and a Support Vector Machine (SVM) model—are provided to the hybrid algorithm. The function generates a hybrid prediction that combines the individual forecasts after applying each model to predict the outcome variable. All three models' predictions are taken into account when creating the hybrid prediction.The hybrid prediction is set to the forecast if all three models concur on it.The hybrid prediction is set to the forecast given by the two models that agree if only two of the models are in agreement. The function uses the probability estimates from all three models to arrive at the final conclusion if no two models agree.

D. MODEL TESTING:

TABLE 1. TABLE FOR MODEL TESTING

PREDICTION	TRUE/FALSE
True positive	True
True negative	True
False positive (Type 1 Error)	False
False negative (Type 2 Error)	False

In this module we test the trained machine learning model using the test dataset.

TABLE 2. FORMULA

Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$
Precision	$\frac{TP}{TP+FP}$

The most commonly used metric to judge a model and is actually not a clear indicator of the performance. The worse happens when classes are imbalanced.Percentage of incidents that are positive out of all the projected instances that are positive.

Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	Positive Predictive Value	$a/(a+b)$
	Negative	c	d	Negative Predictive Value	$d/(c+d)$
		Sensitivity	Specificity		
		$a/(a+c)$	$d/(b+d)$	Accuracy = $(a+d)/(a+b+c+d)$	

Fig 3: Confusion matrix

E. FLASK FRAMEWORK PREDICTION:

The part of a website that a user first interacts with is called the front end. Everything that users view and can interact with is contained in it, including text colours and styles, photos, videos, graphs, and tables, as well as the navigation menu, buttons, and button colours. The front end is created using JavaScript, HTML, and CSS. Python is used to create web apps with Flask. Importing the Flask class was the first step. A new instance of this class is then created. The application's module or package name, '__name__', is supplied as a parameter. To know where to look for resources like templates and static files, Flask needs this information. Flask is then informed which URL should call our method using the route() decorator. The message that should be displayed in the user's browser is returned by this method. Flask is a lightweight Python web framework that makes it simple and quick to create web apps. The Model-View-Controller (MVC) architecture pattern underlies how Flask operates. In this pattern, the model represents the data, the view the user interface, and the controller serves as a bridge between both.

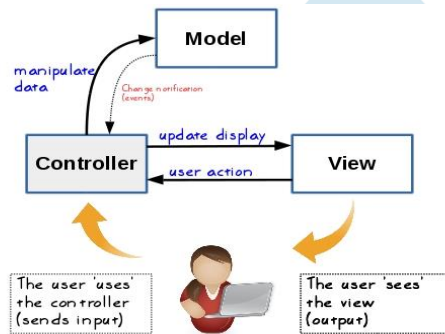


Fig 4: Model View Controller Architecture

F. INSTALL FLASK:

About routes: In your web application, routes are the URLs that users can access. Using the @app.route() decorator and the URL pattern as a parameter, you may construct routes in Flask.

G. CODING:

The system moves onto the coding and testing phase after the design portion is completed. The real system is put into operation during the coding phase by translating the system's design into code written in a specific programming language. Therefore, whenever changes are needed, solid coding practises must be used so that the system can be quickly screwed.

H. TESTING TECHNIQUES:

A programme is tested by being run with the goal of identifying any errors. The implementation stage known as system testing aims to confirm that the system performs accurately and efficiently as anticipated prior to the start of live operation. **Software testing** is a crucial component of software quality assurance because it is necessary for correcting problems. **White box testing** is a technique for creating test cases that derives test cases from the procedural design's control structure. Basis route testing is a white box testing technique that makes use of graph matrices control, flow graph notation, and cyclometric complexity.

Black box testing is the procedure used to confirm that a product's internal operation operates in accordance with specifications and that all internal components have been sufficiently stressed. It makes use of comparison testing, boundary value analysis, equivalence partitioning, and graph-based testing techniques. While **security testing** aims to validate the defences put into a system, condition testing tests the logical conditions present in a module. When testing security, the tester assumes the position of someone who wants to break into the system.

Software must be validated through testing to ensure that it meets user expectations. It is accomplished via a number of black box tests that show requirement adherence. One of two things can happen once a **validation test** is completed: either the function or performance characteristics confirm to specifications and are accepted, or a validation from specification is discovered and a defect is formed. Any system's success depends on how well it is received by its users, and this is tested by regularly staying in touch with them as it is being developed and implementing modifications as needed.

IV.RESULT

The table displays the performance metrics for four different machine learning models, namely Decision tree, Random forest, Support vector machine, and Hybrid, in detecting heart disease, where these models were assessed based on their accuracy score, precision, recall, confusion matrix, and F1 score.

Accuracy Score: The accuracy score shows how many of the model's predictions were accurate. The table shows that the Hybrid model, which scored 95.08% accuracy, was the most accurate, followed by Random Forest (84.62%) and Support Vector Machine (83.08% accuracy). The decision tree model's accuracy score of 78.46% was the lowest.

TABLE 3. PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS ON CLASSIFICATION TASK

ML Algorithms	Accuracy score (in %)	Precision (in %)	Recall (in %)	Confusion Matrix	F1 score (in%)
Decision tree	78.46	86.49	78.05	[[19 5] [9 32]]	
Random forest	84.62	91.89	82.93	[[21 7] [3 34]]	
Support vector machine	83.08	94.59	79.55	[[19 9] [2 35]]	
Hybrid	95.08	90.62	1.0		95.05

Precision: Out of all the instances that were positive identified by the model, precision is the proportion of accurately predicted positive cases. The Support vector machine, Random forest, and Decision tree all had higher precision ratings than the Hybrid model (90.62%, 94.59%, and 91.89%, respectively).

Recall: The proportion of positive cases that were correctly predicted out of all the actual positive cases is known as recall. Random forest came in second with 82.93%, followed by Support vector machine with 79.55%, Decision tree with 78.05%, and the Hybrid model with a flawless recall score of 100%.

Confusion Matrix: The model's actual and predicted classes are displayed in the confusion matrix. The columns reflect the expected classes, whereas the rows represent the actual classes. The confusion matrices demonstrate that the hybrid model accurately predicted all of the positive cases and had the fewest erroneous negatives and false positives. The decision tree model, on the other hand, had the most erroneous positives and negatives.

F1 Score: The F1 score is the harmonic mean of recall and accuracy and offers a general assessment of the model's performance. The Hybrid model, which received an F1 score of 95.05%, came in first, followed by Random Forest (84.62%), Support Vector Machine (83.08%), and Decision Tree (78.46%).

The Hybrid model, which had the highest accuracy, precision, recall, and F1 score, finished with the best overall performance. The model to choose, however, depends on a variety of parameters, including the size of the dataset, the number of characteristics, and the available computational resources.

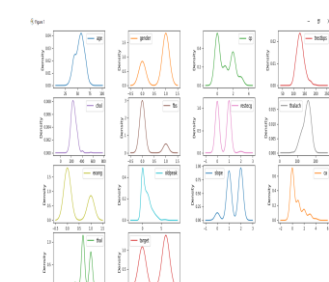


Fig 7: Threshold graph

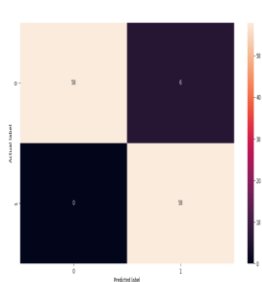


Fig 8: Confusion Matrix of

Hybrid model

The machine learning model used to predict cardiac disease is the main focus of this project. Following data collection, we carry out model training and evaluation, as well as data pre-processing and analysis. A density map and a correlation map are used to display the dataset. A visual depiction of the distribution of every feature in the dataset is provided by the density map, and information about the relationships between the various features is provided by the correlation map. On the categorical attributes during preprocessing, one-hot encoding is done. Following one-hot encoding, the category features are removed from the data, and the remaining characteristics are then normalised with the StandardScaler. Finally hybrid algorithm and flask framework prediction implemented.

V. CONCLUSION

Millions of individuals all around the world suffer from the terrible medical illness known as heart disease. In the early diagnosis and treatment of cardiac disease, machine learning has demonstrated considerable promise. In this work, we investigated a hybrid algorithm's potential for identifying heart illness and discovered that it performed more accurately than other algorithms.

This Study shows how hybrid algorithms have the ability to accurately diagnose heart disease. We can increase precision and reduce errors by combining the benefits of various algorithms. In the healthcare industry, where misdiagnosis can have devastating repercussions, this is especially crucial. The findings of this study have significant ramifications for how heart disease will be diagnosed and treated in the future. Hybrid algorithms have the power to transform how we treat cardiac disease and enhance patient outcomes. Early accurate diagnosis of cardiac disease allows us to take action to stop its progression and enhance patient outcomes.

In conclusion, the application of hybrid algorithms for the detection of cardiac disease shows significant promise. Although these algorithms still need to be improved and refined, the findings of this study lay a solid groundwork for further research in this area. In the end, the application of hybrid algorithms has the potential to reduce mortality and raise the standard of living for people with heart disease.

REFERENCES

1. Wankhede, J.P., Palaniappan, S., Magesh Kumar, S Heart disease prediction using hybrid random forest model integrated with linear model. *Advances in Parallel Computing Int.* vol. 39, pp. 370–376, 2021.
2. A. U. Haq, J. P. Li, J. Khan, M. H. Memon, S. Nazir, S. Ahmad, G. A. Khan, and A. Ali, "Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data," *Sensors*, vol. 20, no. 9, p. 2649, May 2020
3. A. U. Haq, J. Li, M. H. Memon, M. H. Memon, J. Khan, and S. M. Marium, "Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selection," in *Proc. IEEE 5th Int. Conf. Conver. Technol. (ICT)*, Mar. 2019, pp. 1–4
4. U. Haq, J. Li, M. H. Memon, J. Khan, and S. U. Din, "A novel integrated diagnosis method for breast cancer detection," *J. Intell. Fuzzy Syst.*, vol. 38, no. 2, pp. 2383–2398, 2020.
5. Attia, Z., Karras, B., & Kheir, M. (2020). Heart disease diagnosis using machine learning techniques: a review. *Journal of Ambient Intelligence and Humanized Computing*, 11(7), 2953-2964.
6. Pablico, L. S., Su, H. Y., & Chen, K. Y. (2020). A Review on Machine Learning Techniques for Heart Disease Diagnosis. *Journal of Medical Systems*, 44(9), 176.
7. Ribeiro, A. H., Ribeiro, M. H., Paixão, G. M. M., Oliveira, D. M., & Ribeiro, A. F. (2019). Machine learning models for predicting acute myocardial infarction. *Expert Systems with Applications*, 124, 152-159.
8. Pathak, Y., Kumar, M., & Bhatia, V. (2021). Comparative Analysis of Machine Learning Techniques for Heart Disease Diagnosis. *International Journal of Intelligent Systems and Applications*, 13(2), 40-48.
9. Liu, F., Zhang, Q., & Yang, F. (2019). Application of machine learning algorithms in diagnosis of heart disease. *Journal of Healthcare Engineering*, 2019, 1-11.
10. Abdi, A., Ahmadi, M., & Alikhani, M. (2021). Machine learning-based approach for predicting the risk of heart disease. *Journal of Ambient Intelligence and Humanized Computing*, 12(6), 6149-6160.
11. Assegie, T.A., Rangarajan, P.K., Kumar, N.K., Vigneswari, D. An empirical study on machine learning algorithms for heart disease prediction. *IAES International Journal of Artificial Intelligence*, vol 11, pp. 1066-1073, 2022
12. Liu, X., Xie, J., Yang, X., & Deng, W. (2021). A novel feature selection and classification method for heart disease detection based on machine learning. *Computer Methods and Programs in Biomedicine*, 204, 106078.