WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Heart Disease Detection Using Machine Learning Techniques

Faizan Younas[a,*], Saqib Ubaid[b], Ali Raza[c], Ijaz Azmat[d], Farhan Akhtar[e]

[a,b,c]*Department of Computer Science, Khwaja Fareed University of Engineering & Information Technology, Rahim Yar Khan, 64200, Pakistan.*
[d]*Faculty of Computer Science, Ilma University, Korangi Creek, 75190, Pakistan.*
[e]*Department of Mathematics, Khwaja Fareed University of Engineering & Information Technology, Rahim Yar Khan, 64200, Pakistan.*
DOI: https://doi.org/10.55248/gengpi.2022.3.2.15

A B S T R A C T

According to a recent World Health Organization survey, 17.9 million people die each year as a result of heart-related disorders, and the number is continuously growing. With a rising population and sickness, it is becoming more difficult to diagnose disease and provide proper therapy at the appropriate time. However, there is a ray of optimism in that recent technological breakthroughs have expedited the public health sector by generating sophisticated functional biomedical solutions.The purpose of this paper is to examine the different data mining techniques, namely Naive Bayes, Random Forest Classification, Decision Tree, and Support Vector Machine, using a qualified dataset for Heart disease prediction, which includes attributes such as gender, age, chest pain type, blood pressure, blood sugar, and so on. The research entails determining the correlations between the different qualities of the dataset using typical data mining techniques and then using the attributes to forecast the likelihood of a cardiac condition. These machine learning algorithms need less time to forecast disease with more precision, reducing the loss of priceless lives all across the world.

Keywords:Heart Disease, Data Mining, Machine Learning.

1. Introduction

Computers are employed for the inspection of internal organs of the physical structure, illness diagnostics, advanced computer-based systems are used to analyze sensitive organs of the body, and certain difficult procedures may be performed with the assistance of computers. Magnetic Resonance Imaging (MRI) is used by doctors to check the status of soft tissue within the body. MRIs employ intense magnetic fields to scan and recreate an image of the interior body structure. The MRI scanner transmits activity sensed inside back to the computer, which recreates an image of a specific location. The computer can build a three-dimensional representation of activity within the region being scanned, allowing clinicians to examine the activity within a patient's body and correctly diagnose them. Without computers, doctors would have to undergo invasive surgery to uncover physical and operational flaws within the body to make a diagnosis.

Machine learning is rapidly entering many aspects of the healthcare business, from diagnosis and prognosis to drug discovery and epidemiology, and has the potential to significantly revolutionize the medical environment. Disease recognition and diagnosis are at the forefront of ML research in medicine. Clinical researchers may "reconstruct the basic processes of illness" by combining machine learning with multimodal information and nearly infinite computer capacity.

The heart is one of the most important organs in the human body. It is a muscular organ that pumps blood into the body and is the heart of the cardiovascular system. The cardiovascular system is made up of all blood vessels, including arteries, veins, and capillaries, that form a complicated network of blood circulation throughout the body. Any restriction or anomaly in normal blood circulation or flow from the heart might result in a variety of serious heart disease problems. These are known as cardiovascular diseases (CVDs) and are among the world's worst diseases.

* *Corresponding author.*
E-mail address: alirazaatofficial@gmail.com

Data has proved to be the fuel for health in this age of technology and digitalization. Almost all medicine and clinics institutes now keep their patients' data in an electronic format. This covers their medical records, symptoms, diagnosis, length of sickness, recurrences, and any fatalities. As a result, the amount of medical data created daily is steadily expanding. If the available data is used to construct screening and diagnostic models, it would not only relieve the load on medical workers but will also help in early diagnosis and timely treatment for patients, therefore significantly improving the health system. Furthermore, it can help in the development of a monitoring and preventative program for people who, based on their medical history, are at risk of developing CHD. It is associated with numerous risk factors in heart disease and a necessity of the time to obtain accurate, trustworthy, and reasonable techniques to establish an early diagnosis to accomplish fast disease treatment. In the healthcare arena, data mining is a widely utilized approach for processing massive amounts of data.

Predictive modeling is a method for creating a model that can generate predictions. This model comprises a machine learning method, which allows data-driven models to learn particular information from observed data in a training data set to generate those predictions. In this study, a specific goal is employed to estimate output for new use cases, which indicates whether or not the cardiac disease is present. As a result, a supervised learning classification method would be ideal for training the data.

The most difficult aspect of cardiac disease is detecting it. Some tools can forecast heart disease, but they are either too expensive or too inefficient to quantify the risk of heart disease in humans. Early identification of heart disorders has been shown to reduce mortality and overall consequences. However, it is not possible to correctly monitor patients every day in all circumstances, and consultation with a doctor for 24 hours is not available since it takes more intelligence, time, and competence. We may use various machine learning algorithms to analyze data for hidden patterns in today's environment since we have a lot of data. Hidden patterns in medical data can be utilized for health diagnosis.

2. Research contributions

Our goal is to predict if people have heart disease based on user characteristics. This is significant in the medical industry. If such a forecast is sufficiently accurate, we may not only prevent incorrect diagnosis but also save human resources. When a patient who does not have heart disease is diagnosed with heart disease, he will experience unneeded anxiety, and when a patient who does have heart disease is not diagnosed with heart illness, he will lose out on the best opportunity to treat his disease. Such misdiagnosis is upsetting for both people and hospitals. We may avoid unneeded difficulties by making precise predictions. Furthermore, if we can apply our machine learning technique to medical prediction, we will save human resources by eliminating the need for complex diagnosis processes in hospitals. The main objective of developing this project is:

- To develop a machine learning model to predict the future possibility of heart disease by implementing Machine learning Algorithms.
- To determine significant risk factors based on a medical dataset that may lead to heart disease.
- To analyze feature selection methods and understand their working principle.

3. Research Scope

The purpose of this study is to identify the most significant risk factors for cardiovascular diseases (CVDs) in patients and predict IHD using various classification algorithms. Overall, the goals of this thesis may be summed up as follows:

- To identify the most significant risk factors of CVDs.
- To find the correlation between CVDs and risk factors.
- To predict cardiovascular diseases (CVDs) using mostly significant risk factors by applying different classification algorithms.

Our research goal is to determine whether or not patients have heart disease based on a variety of patient characteristics. Our project's objective is to conserve human resources in medical centers while improving diagnosis accuracy. In our initiative, we employ a variety of ways to diagnose cardiac illness.

4. Advantages of Proposed System

The suggested technology functions as a decision support system and will assist clinicians with diagnosis. One of the key benefits of machine learning is the detection and diagnosis of diseases and conditions that would otherwise be difficult to identify. We can deliver better information to clinicians at the moment of patient care if we use this advanced sort of analysis. When we meet patients regularly, we have simple access to their blood pressure and other vital indicators. We need to provide more information to doctors so that they can make better judgments regarding patient diagnoses and treatment options, while also knowing the potential consequences and costs associated with each. The value of machine learning use cases in health insurance is its ability to process large datasets beyond the scope of human functionality, and then reliably transform the analysis of data into clinical insights that help doctors plan and provide care, resulting in better outcomes, lower prices than attention, and increased patient satisfaction. The primary goal of this research project is to create an expert system for heart disease detection based on advanced machine learning algorithms. Compare its performance to that of other computer-based decision assistance systems that are now available. The supervised model's predicted accuracy and training time are greatly improved. The model successfully predicts the severity of cardiac disease in patients.

5. Literature Review

A variety of prediction systems for various diseases have been suggested and executed using various methodologies. Previous research on heart disease by many authors investigated tested and analyzed various ways. They considered the data set from the UCI data repository, which can also be obtained from Kaggle, for the work's implementation. The categorization and prediction approach was applied to the data set by the authors. The authors are Sellappan Palaniappan and Ra ah Awang, and they employed three data mining classification modeling strategies. These methods extract hidden information from the heart disease database (Alotaibi, 2019). They utilized the DMX query language to gain access to the model.

To assess the outcomes, the model is trained on train data and evaluated on test data. The Lift chart and classification matrix approach was employed by these authors to assess the model's efficacy. The algorithm mines the historical heart disease database for hidden insights. The authors built the suggested system using the.net framework, which is a web-based prediction system that is simple to use, reliable, extendable, and scalable. The Naive Bayes model is the most successful for predicting cardiac disease, followed by the Decision Tree. The Decision Tree was outperformed (Bommadevara, (2019).

The researchers of the heart disease prediction system are Mai Showman, Tim Turner, and Rob Stocker, who used the K-Means approach in conjunction with the Decision Tree. To improve model accuracy, they used initial centroid selection strategies in the execution of this study. They employed the Decision Tree classifier to diagnose heart problems. There are a variety of prediction systems developed for different illnesses and implemented using different ways to produce the first centroids depending on the number of real samples in the data set the Random Row.

Previous research on heart disease by many authors investigated tested and analyzed various ways. They considered the data set from the UCI data repository, which can also be obtained from Kaggle, for the work's implementation. The categorization and prediction approach was applied to the data set by the authors. The precision achieved by Naive Bayes is 84.15 %, whereas Random Forest achieved 84.16 percent, for which it obtained almost comparable precision when applied with selected highlights for both techniques (Bommadevara, (2019).

The researchers who used machine learning algorithms for cardiac disease prediction include Bommadevara, H. S. A., Sowmya. Y and Pradeepini, G (Zafar1). Praveena, M., Asha Deepika, R., and Sai Raghavendhar researched the prediction of heart disease using data mining approaches (Alotaibi, 2019). Sai Krishnasree, K., and M. R. Narasinga Rao conducted comparison research on Bayesian regularisation with multiple regression models (Pahwa) Siva Kumar, P., Anand, D., Uday Kumar, V., and Bhattacharyya conducted a study on a computational knowledge technique for the feasible detection of cardiac disease using a genetic algorithm (Srinivas, (2018).

6. Proposed Methodology

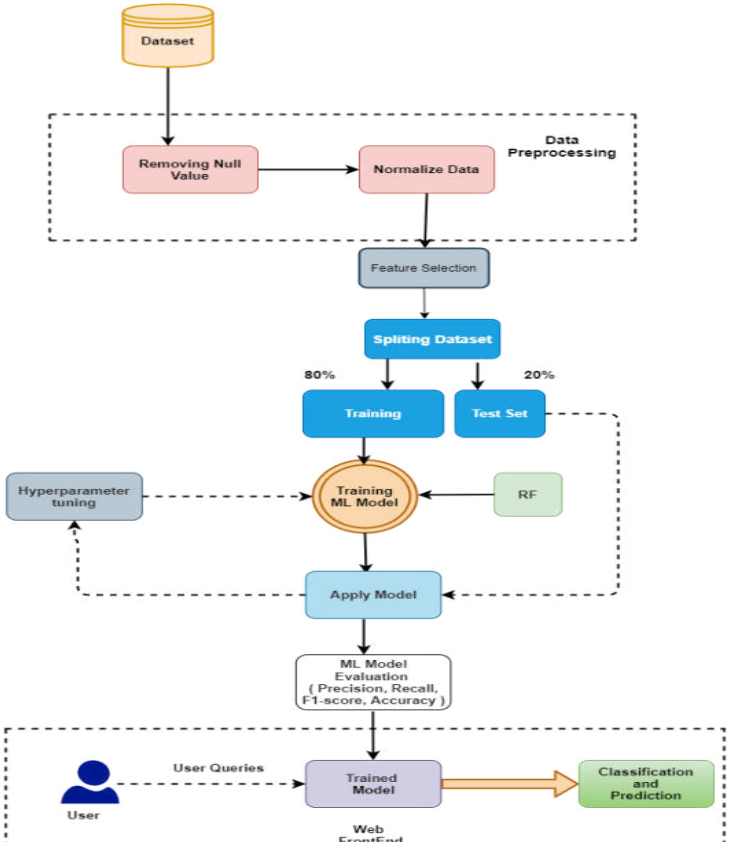

Fig. 1 –proposed system methodology.

The System methodology Diagram, assist us to identify the building pieces of the proposed system, without having to determine, yet, if they are to be Hardware, Software, or mechanical. The heart disease dataset is utilized in this research study. To get accurate results from the proposed model dataset preprocessing is applied. The preprocessing involves unnecessary noise cleaning, data transformation, and data normalization. The TFIDF is utilized for feature engineering. The dataset splitting is conducted in the next phase. The 80% portion of the dataset is used for the proposed model and 20% is used for model testing and result evaluation. The proposed random forest machine learning approach is fully hyperparameter tuned. The best fit model is deployed behind the heart disease website.

7. Dataset

Kaggle provided us with the data set that we used in our study (https:// www.kaggle.com/ronitf/heart-disease-uci). The dataset that we utilized in our study comprises 14 columns and 303 rows in total. The first 13 columns are the features that we will use later to forecast the last column, 'diagnostic,' which will tell us whether or not the patient has heart disease. The 303 rows reflect information from 303 patients discovered in the collection. The Cleveland Heart Disease dataset from the UCI repository was utilized in this experiment (Yasin, 2019). In that vein (pun intended), the dataset below contains 303 observations, 13 features, and 1 target attribute. The findings of the aforementioned non-invasive diagnostic tests, as well as other pertinent patient information, are among the 13 characteristics. The target variable comprises the invasive coronary angiography result, which shows the presence or absence of coronary artery disease in the patient, with 0 indicating no CHD and labels 1-4 denoting CHD. The majority of research using this dataset has focused on attempting to distinguish presence (values 1,2,3,4) from absence (values 1,2,3,4). (Value 0).

7.1. Dataset Attributes

The dataset consists of 303 individuals' data. There are 14 columns in the dataset, which are described below.

- ***Age:*** displays the age of the individual.

- ***Sex***: displays the gender of the individual using the following format:
  1 = male
  0 = female

- ***Chest-pain type***: displays the type of chestpain experienced by the individual using the following format:
  1 = typical angina
  2 = atypical angina
  3 = non—anginal pain
  4 = asymptotic

- ***Resting Blood Pressure***: displays the resting blood pressure value of an individual in mmHg (unit)

- ***Serum Cholesterol***: displays the serum cholesterol in mg/dl (unit)

- ***Fasting Blood Sugar***: compares the fasting blood sugar value of an individual with 120mg/dl.
  If fasting blood sugar > 120mg/dl then : 1 (true)
  else : 0 (false)

- ***Resting ECG:*** displays resting electrocardiographic results
  0 = normal
  1 = having ST-T wave abnormality
  2 = left ventricular hyperthrophy

- ***Max heart rate achieved:*** displays the max heart rate achieved by an individual.

- ***Exercise-induced angina:***
  1 = yes
  0 = no

- ***ST depression induced by exercise relative to rest:*** displays the value which is an integer or floats.

- ***Peak exercise ST segment:***
  1 = upsloping
  2 = flat
  3 = downsloping

- **Several *major vessels (0–3) colored by fluoroscopy:*** displays the value as integer or float.

- ***Thal:*** displays the thalassemia:
  3 = normal
  6 = fixed defect
  7 = reversible defect

- ***Diagnosis of heart disease:*** Displays whether the individual is suffering from heart disease or not:
  0 = absence
  1, 2, 3, 4 = present.

8. Correlation analysis

The correlation matrix between those features is the best technique to compare the relationship between those features. A correlation matrix is a table that displays the coefficients of correlation between variables. Each cell in the table represents the relationship between two variables. A correlation matrix can be used to summarise data, as an input for more advanced analysis, or as a diagnostic for further studies.

Fig. 2–dataset correlation analysis.

9. Data Preprocessing

The goal of data preprocessing is to extract valuable data from raw heart disease datasets and then these data should be translated into the format essential for the prediction of the risk level. Because of the useless information in the heart disease datasets, the original raw data cannot be utilized directly in the prediction approach; hence, raw data must be cleaned, evaluated, and converted for future stages during the data preparation phase. As a result, the technique is used to identify the unnecessary characteristics, which are then translated into the row-column format after the irrelevant ones are removed. Each row provides the patient's information, while each column represents a set of qualities (biomarkers). The last column displays the class designation that corresponds to the heart patients' risk level. Data preprocessing is a data mining approach that entails converting raw data into a usable format. Real-world data is frequently inadequate, inconsistent, and/or deficient in specific behaviors or patterns, and it is rife with inaccuracies. Data preparation is a tried-and-true way for overcoming such problems. Machine Learning Methodology Import the libraries, Import the data-set, and Check for missing values are all steps in Data Preprocessing.

10. Features engineering

In this part, we will distribute the goal value, which is critical for selecting acceptable accuracy metrics and, as a result, accurately assessing different machine learning models. First and foremost, we will count the values of the explanatory variable, also known as the deciding variable, which will tell us if a patient has heart disease or not. First and foremost, we will distinguish between numerical and category characteristics. Then we'll plot the relationship between the category features and try to find out or rather notice the effect of those categorical features on the real deciding variable "diagnosis."

11. Data Splitting

The process of dividing accessible data into two pieces, generally for cross-validation reasons, is known as data splitting. One component of the data is used to create a predictive model, while the other is utilized to assess the model's performance. The training set is used to create models and feature sets; it serves as the foundation for estimating parameters, comparing models, and all of the other actions necessary to arrive at a final model. The test set is only used at the end of these operations to estimate a final, unbiased evaluation of the model's performance. Looking at the results of the test sets would skew the findings since the testing data would have been part of the model-building process. In any Machine Learning model, we will divide the data into two independent sets: Training Set and Test Set. In general, we divide the data set into 70:30 or 80:20 ratios, which indicates that 70 percent of the data is taken in train and 30 percent is taken in the test. However, the Splitting might vary depending on the form and size of the data set. In our example, we divided the data set 80:20. In this part, we will go over the dataset.

12. Proposed Machine Learning Classifier

The overall purpose of the study is to anticipate the incidence of heart disease with the greatest precision possible. To do this, we will put numerous categorization algorithms to the test. This section summarizes all of the study's findings and introduces the best performance based on the accuracy metric. Throughout classification approaches, we picked many algorithms that are commonly used to solve supervised learning challenges. First of all, let's equip ourselves with a useful tool that benefits from the coherence of the SciKit Learn library and design a generic function for training our models. First and foremost, let us provide ourselves with a useful tool that makes use of the coherence of the SciKit Learn library and develops a generic function for training our models. The objective for reporting accuracy on both the train and test sets is to let us determine if the model over or under fits the data (the so-called bias/variance tradeoff). The data will then be divided, tested, and trained in an 80:20 ratio. Then we'll build a model in which we'll run all of our algorithms.

Random Forest methods are employed in both classification and regression. It builds a tree for the data and then makes predictions based on it. The Random Forest approach may be used to big datasets and generate consistent results even when substantial sets of record values are missing. The decision tree samples that are created can be stored and used on additional data. There are two steps in the random forest: first, generate a random forest, and then make a prediction using a random forest classifier established in the first stage.

$$RFfi_i = \frac{\sum_j norm\ fi_{ij}}{\sum_{j \in all\ feature\ ,k \in all\ tress} norm\ fi_{ik}} \qquad (1)$$

13. Development

Based on current and predicted user demands, this procedure supports existing infrastructure requirements and makes specific suggestions for hardware and network solutions. Application requirements, data resources, and organizational people all play a role in identifying the best hardware solution. It is illustrated by a three-tier design that includes a graphical interface, process improvement, and database management. It depicts the system's parts, the services they offer, and how they communicate for the system to work.

13.1. Tool Selection

- Jupyter Notebook
- Web browser for testing.
- Draw.io for diagram.
- Visual Studio Code
- Microsoft word.

13.2. Hardware

- PC with processor P4 and above
- RAM at least 2GB

The development phase is addressed in this section. Model and web page coding are described. In addition, user interface samples from the website are displayed to provide a clear project view. This phase indicates that the project is finished and ready for testing and implementation.

14. Proposed System Testing

The work is created, executed, and tested in stages (a bit more is added each time) until the product is complete. It involves both development and maintenance. When a product meets all of its specifications, it is said to be done. First, we take a dataset and divide it into two parts: train and test, and then we use machine learning to train multiple models on the training data. After finishing that process, we evaluated the model on the testing data to see if our model was overfitted, under fitted, or worked correctly. The front-end is built in the second phase using HTML, CSS, and Django. After developing the element, we tested the application to ensure that it is responding to the requests and can be correctly modified. We tested the final component after finishing that work. All of these phases were constructed one at a time, and each module was added one at a time.

Table 1–Test case analysis.

Parameter	Value
TestCase ID	01
ApplicationName	Heart Disease Detection using ML
Purpose	Todescribetheapplicationformin whichusersign up
Environment	Python (Django)
Pre-Requisite:	The userwillbetheperson
Strategy:	The userwillperformthefollowingoperation Provideinformation. Saveinformation.
ExpectedResult	UsersshouldentertheLoginformwiththeirrelevantprivileges.
Observations	The user has successfully entered the system havingdefined by the admin
TestCase ID	02
ApplicationName	Heart Disease Detection using ML
Purpose	Todescribetheapplicationforminwhichcustomerslog in
Environment	Python Django
Pre-Requisite:	The userwillbethe person
Strategy:	The userwillperformthefollowingoperation Enternameand password.
Expected Result	Users should see their Prediction Requirements form successfully, if any incorrect information the error occurs.
Observations	The user has successfully entered the system having defined by the admin
TestCase ID	03
ApplicationName	Heart Disease Detection using ML
Purpose	Todescribetheprediction forminwhichthe user enters the value
Environment	Python Django
Pre-Requisite:	The userwillbetheperson
Strategy:	The userwillperformthefollowingoperation Enter all required values.
ExpectedResult	UsersshouldentertheresultingformwiththeirrelevantprivileGes.
Observations	The user has successfully entered the system havingdefined bythe admin
TestCase ID	04
ApplicationName	EagleList
Purpose	Todescribetheapplicationforminwhich the adminviewclient
Environment	Python Django
Pre-Requisite:	The userwillbetheadmin
Strategy:	The userwillperformthefollowingoperation Entere-mail,password. Viewclient.
ExpectedResult	UsersshouldenterthemainformwiththeirrelevantprivileGes.
Observations	The user has successfully entered the system having defined by the admin

In this section, we wrap up all of the project's testing and development work. It is checked in several ways to ensure that there are no errors. Whether or not the system is operational. Test cases also provide explanations.

15. Result and Evaluation

Using the Cleveland training data set with 14 distinct features, several experiments are carried out to test the performance and validation of the generated model. The results are estimated using confusion matrix measurements, and the accuracy is compared using several approaches. Using the confusion matrix, several parameters such as sensitivity, specificity, accuracy, and precision may be derived to assess the prediction model's validity. Specificity is the fraction of properly-recognized negatives, whereas sensitivity is the proportion of correctly detected true positives. The following formulae can be

used to express these measurements quantitatively. Where TP, TN, FP, and FN represent True Positive (number of positive data correctly labeled by the classifier), True Negative (number of negative data correctly labeled by the classifier), False Positive (number of negative data incorrectly labeled as positive), and False Negative (number of positive data incorrectly labeled as negative).

- Accuracy (all **correct** / all) = TP + TN / TP + TN + FP + FN
- Misclassification (all **incorrect** / all) = FP + FN / TP + TN + FP + FN
- Precision (**true** positives / **predicted** positives) = TP / TP + FP
- Recall (**true** positives / all **actual** positives) = TP / TP + FN

For the first experiment, we employed the train-test split validation approach, in which our data set was divided into two halves, and we performed numerous tests with varying percentages; the optimum splitting we got was 80 percent of the data for training and 20% for testing. The outcomes are produced by employing RF algorithms. Based on the testing findings, it is obvious that the RF algorithm has the best classification accuracy. However, the train-test split validation approach frequently results in overfitting since the assessment is mostly dependent on which data is used in the training set and which data is used in the test set. As a result, the evaluation may change greatly depending on how the split is formed. As a result, we introduced the cross-validation approach to address this issue. Cross-validation is a strong anti-overfitting strategy that leverages the original training data to produce several small train test splits to modify the model. The chart shows that the RF still performed better, creating the model with an accuracy of 90.16 percent. We may infer that RF outperforms and outperforms the other algorithm methods in terms of performance and accuracy.

Table 2–Proposed approach results evaluation.

Category no.	Precision-score	Recall-score	F1-score	Support-score
0	0.93	0.79	0.85	33
1	0.85	0.95	0.90	43

A Confusion matrix is a N x N matrix that is used to assess the effectiveness of a classification model, where N is the number of target classes. The matrix compares the actual goal values to the machine learning model's predictions. This provides us with a comprehensive picture of how well our classification model is working and the kind of errors it is producing.

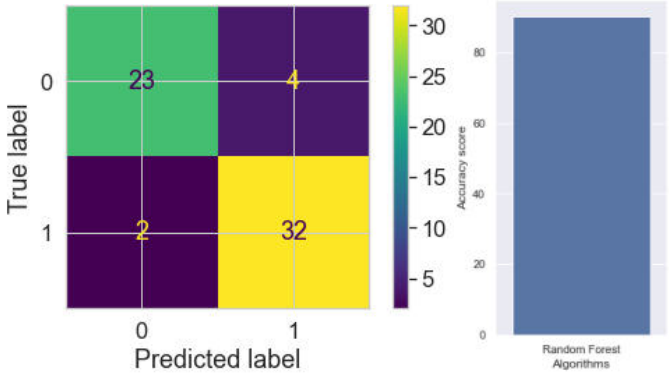

Fig. 3–(a)confusion matrix, (b) proposed model accuracy score.

16. Conclusion

In this study, we apply the random forest method to the data set, assess the outcomes, confirm their correctness, and plot the confusion matrix. In addition, consider the true positive, true negative, false positive, and false negative values. Finally, we decide that our program will be the greatest and will function effectively. We put forth a lot of effort and achieved fantastic results. In the future, we will focus on additional diseases such as renal illness and breast cancer. And the user can forecast the likelihood of this sickness occurring.

Acknowledgements

We gratefully acknowledge the support of our institute KFUEIT and our research supervisors for their support and appreciation. We would like to thank everyone who had contributed to the successful completion of this research.

REFERENCES

Alotaibi, F. S. (2019). Implementation of Machine Learning Model to Predict Heart Failure Disease",. *(IJACSA) International Journal of Advanced Computer Science and Applications, 10*.

Bommadevara, H. S. ((2019). *Heart disease prediction using machine learning algorithms.* IEEE.

Pahwa, K. R. (n.d.). *Prediction of heart diseaseusing hybrid technique for selecting features.*

Srinivas, V. A. ((2018). *A novel approach for Heart disease prediction: Machine learning techniques.*

Yasin, S. A. (2019). *Analysis of single and data mining techniques for heart disease prediction using real time dataset.*

Zafar1, S. S. (n.d.). *A review on heart diagnosis based on decision support systems:.*