Journal of Computer and Communications, 2024, 12, 134-150
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

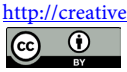
Identification of Cardiovascular Disease via Diverse Machine Learning Methods

Araf Islam[1*], Mohammad Abu Saleh[2], Afia Fairooz Tasnim[3], Md. Samiun[2], Syeda Kamari Noor[4], Kanchon Kumar Bishnu[5]

[1]Department of Computer Science, Westcliff University, Irvine, USA
[2]Department of Business Administration, International American University, Los Angeles, USA
[3]Department of Public Health, California State University, Los Angeles, USA
[4]Department of Business Administration, Westcliff University, Irvine, USA
[5]Department of Computer Science, California State University, Los Angeles, USA
Email: *a.islam.585@westcliff.edu

How to cite this paper: Islam, A., Saleh, M.A., Tasnim, A.F., Samiun, Md., Noor, S.K. and Bishnu, K.K. (2024) Identification of Cardiovascular Disease via Diverse Machine Learning Methods. *Journal of Computer and Communications*, **12**, 134-150.
https://doi.org/10.4236/jcc.2024.1212009

Received: October 23, 2024
Accepted: December 23, 2024
Published: December 26, 2024



Abstract

Over the past ten years, there has been an increase in cardiovascular disease, one of the most dangerous types of disease. However, cardiovascular detection is a technique that analyzes data and precisely diagnoses cardiovascular disease using machine learning algorithms. Early diagnosis may lead to better outcomes for heart treatment. Then, utilizing machine learning to detect cardiac disease will be easy in a couple of seconds. This study proposes an automatic way for detecting cardiovascular diseases such as heart disease using machine learning. A physician's accurate and thorough evaluation of a patient's cardiovascular risk plays a critical role in lowering the incidence and severity of heart attacks and strokes as well as improving cardiovascular protection. To develop technology for the early detection of cardiovascular disease, the Kaggle dataset was gathered. Certain preprocessing techniques were used to improve accuracy and outcomes. Ultimately, we employed decision trees, logistic regression, and random forests to reach our objective. Of these, random forest yielded the highest accuracy of 96%, making them useful for obtaining high-quality results with greater precision.

Keywords

Decision Tree, Logistic Regression, Random Forest, Preprocessing, Machine Learning, Accuracy

1. Introduction

Since the heart controls blood circulation, it is the most vital component of the

human body. Many people are suffering from diverse types of heart diseases. Now cardiovascular disease is an extremely dangerous heart disease. Cardiovascular diseases accounted for almost half of all deaths in the US and other countries. There are diverse types of risk factors of heart diseases. Tobacco use is one of them. Physical inactive, and low diet are also affecting. There is a risk factor named air pollution, which is known as an environmental factor. Behavior-related health risks can cause people to have elevated blood sugar levels, have diabetes, elevated levels of cholesterol, and to be overweight. These "intermediate risk factors" can be examined in clinics to identify a higher risk of stroke, heart attack, heart failure, and so on [1]. All primary healthcare facilities must have accessibility to fundamental wellness equipment and medicines for non-communicable illnesses to make sure that those in need acquire medical care and guidance. Many people from different countries are dying because of cardiovascular disease. If this disease can be detected in the early stage, people will be alert and take necessary steps to cure this disease. By identifying those who are suffering from CVDs and guaranteeing that they get proper therapy, early deaths can be declined. Our cardiovascular detection method uses machine learning algorithms to analyze data and reliably identify cardiovascular. The Kaggle dataset used to detect heart disease was obtained. Four databases make up this 1988 data set: Hungary, Cleveland, Long Beach V, and Switzerland [2]. Unfortunately, over the past 40 years, there has been a global increase in cardiovascular cases. Maximum patients who are suffering from CVD are caused by stroke, and they are over 70 years old. Early death is common among those people. When asked to name the major cause of death in the United States, well than half (51%) of participants in a Harris Poll from 2023 did not mention heart disease on behalf of the Heart Association of the United States [3]-[10].

Heart failure (9.1%), high cholesterol levels (13.4%), stroke (17.5%), various major Cardiovascular causes combined (17.1%), chronic heart disease (40.3%), and artery diseases (2.6%), are among the causes of cardiovascular mortality. To ensure that those in need receive medical attention and guidance, basic wellness supplies and medications for non-communicable diseases must be readily available in all primary healthcare institutions. Cardiovascular disease is the leading cause of death for people worldwide. Consequently, the death rate from cardiovascular disease is rising every day. If detected early enough, CVD is curable even though it spreads to other organs like the brain and bones more quickly than other types of cancer. But if it spreads to other body parts, the chances of recovery completely vanish, and the therapies become more challenging. Building an artificial neural network that can function well on challenging tasks requires careful consideration of feature extraction and selection [11]-[19]. They help reduce the noise of the data, delete unwanted data, and enhance the precision of the model. It is crucial for the early diagnosis of cardiovascular disease as a result. For this reason, this technique has developed to quickly identify this disease in the early stage. This study suggests utilizing machine learning to automatically diagnose cardiovascular disorders, including heart disease. Reducing the frequency and severity of heart

attacks and strokes as well as enhancing cardiovascular protection are largely dependent on a doctor's precise and comprehensive assessment of a patient's cardiovascular risk. To improve accuracy, three models were used with different algorithms.

2. Literature Survey

Early fatalities from CVDs may be reduced by identifying those who are more vulnerable and making sure they receive the right care. ensuring that all clinics provide non-communicable disease prescriptions and the barest minimum of healthcare supplies to anyone in need of guidance or treatment [1] [20]-[28]. Inspection for cardiovascular illness requires specialized methods and equipment, which is expensive and time-consuming. It also requires more training and experience to identify this illness. This is the reason that machine learning has been used recently to investigate automatic cardiovascular disease detection. A summary of several recent works on automated cardiovascular disease diagnosis is given in the paragraphs that follow.

For instance, A. Nikam *et al.* [29] trained machine learning models to identify cardiovascular illness using a dataset of 70,000 patient records that was gathered from the healthcare industry. They collected a dataset of 12 rows and 70,000 columns. They have used preprocessing techniques including feature selection, validation, and test set generation in addition to data cleaning to get rid of extraneous data and binary classification using various parameters. To diagnose heart illness, they ultimately used a Decision tree, LGBM, XGB, Neural Network, Logistic Regression, KNN, and Naive Bayes [30]-[37]. In the outcome section, the authors' accuracy in the Decision tree GBDT model was 73%. P. Chinnasamy *et al.* [38] presented a method of CVD detection based on machine learning classification. In this study, the authors suggested an autonomous machine learning-based approach for identifying cardiovascular disease. To determine the results, the authors utilized the dataset which was gathered from Kaggle. Before being uploaded to the system, the data underwent preprocessing. However, the researchers discovered that there was an imbalance in the dataset, with 20% of the data being cardiovascular disease and the remaining 80% normal heart. Lastly, the authors discovered that the accuracy of Logistic regression, SVM, Decision tree, and GBDT classifier was 87%, 87%, 78%, and 82%, respectively. The M. M. Ali and others [39] developed a cardiovascular disease detection system for better self-examination using machine learning. In this experiment, a public dataset was used from Kaggle with 14 attributes, and includes the records of 1025 patients, including 312 women and 713 men; 49% of the patients are healthy, while 51% have cardiac disease. Before applying any model, data preprocessing was applied. The author used Weka for mining data, data cleaning and other techniques. For the highest accuracy, the author used LR, ABM1, MLP, KNN, DT, and RF. The lowest accuracy achieved for LR was only 87%. Finally, KNN provides the highest accuracy of 90%. To improve self-examination, the S. S. Apte, and C. S. Dangare [40]

used machine learning to create a system for detecting cardiovascular illness. A public dataset from the Cleveland Heart Disease database with 303 records was utilized in this experiment. It has 13 input attributes in total, and they use all inputs to take inputs from other patients manually. Data preprocessing was done before any models were used. Different techniques were utilized by the author for data cleaning, mining, and other tasks. The author employed Decision tree, Naive Bayes, and Neural Networks for optimal accuracy. With Naive Bayes, the lowest accuracy of only 87% was attained. Lastly, Neural Networks offer 90% accuracy, which is the highest. A machine learning-based classification technique for CVD detection was presented by Purushottam *et al.* [41].

In this paper, the authors proposed a method for detecting cardiovascular illness based on autonomous machine learning. The authors used the dataset that they had collected from Cleveland Heart Disease database to ascertain the outcomes. Preprocessing was done on the data before it was posted to the system. But the researchers found that the dataset was unbalanced, with 80% of the data representing normal hearts and 20% representing cardiovascular illness. Finally, the scientists found that the Decision tree classifier had an accuracy of 87%. By utilizing machine learning, M. Rana *et al.* [42] created a system for detecting cardiovascular illness to improve self-examination. Using a public dataset from a Kaggle with 14 attributes, the records of 1025 patients, including 312 women and 713 men—were used in this experiment. Of the patients, 52% had cardiac illness and 48% were in good condition. Data preprocessing was done prior to the use of any models. The author employed Weka for several strategies such as data cleansing and mining. The author employed DT, and RF to achieve the maximum accuracy. Only 87% accuracy was the lowest for DT. RF offers the maximum accuracy of 89%, to sum up. K. Battula and others [43] developed a technique for identifying cardiovascular disease using machine learning to enhance availability. This experiment used a publicly available dataset (303 records) from the Cleveland Heart Disease database. There are a total of 13 input qualities, and they use them all to manually collect input from other patients. Prior to using any models, the data was pre-processed. The author employed various methodologies for tasks such as data mining and cleaning. For the best accuracy, the author used KNN, decision trees, and ANN. The lowest accuracy of just 85% was obtained using Naive Bayes. Finally, and most accurately, ANN provides 91% accuracy. The Md. R. Ahmed *et al.* [44] trained machine learning models to identify cardiovascular illness using a dataset of 70,000 patient records that was gathered from the healthcare industry. They collected a dataset of 12 rows and 70,000 columns. They have used preprocessing techniques such as feature selection, validation, and test set generation, along with data cleaning to eliminate superfluous data and multiclass and binary classification. To diagnose heart illness, they used a Decision tree, Neural Network, Logistic Regression, KNN, and Naive Bayes. In the outcome section, the authors' accuracy in the Naive Bayes model was 78%. Sobuz *et al.* [45] innovated the application of SCBA by utilizing machine learning techniques. They have used a

range of machine learning methods, including random forest and artificial neural networks, to predict the novel and mechanical properties of the LWC. With an accuracy of 0.92, the random forest did exceptionally well. Datta *et al.* [46] looked at the mechanical aspects using machine learning approaches. They have used XGBoost and KNN. XGBoost fared better than KNN with an accuracy of 93%. Hasan *et al.* [47] investigated the connection between mechanical characteristics and varying POFA concentrations using state-of-the-art machine learning techniques. They employed random forest, which produced results with an accuracy of 92%.

The automatic detection of cardiovascular disorders, including heart disease, by machine learning is proposed in this work. Improving cardiovascular protection and reducing the frequency and severity of heart attacks and strokes depend heavily on a doctor's precise and comprehensive assessment of a patient's cardiovascular risk. The cardiovascular disease was identified in this study by means of machine learning techniques and data. From the Kaggle dataset, 70,000 patients' records were used. Data mining with different techniques, feature selection, the binary classification, data cleaning, and other approaches were applied in preprocessing to avoid overfitting. Additionally, random forest, logistic regression, and decision tree were applied on the dataset. In conclusion, the accuracy of the random forest model was higher than that of previous studies.

Machine learning is used in this paper to detect cardiovascular disease. This work has made the following notable contributions:

- A major contribution of this project is to apply pre-processing techniques to the collected dataset, which contains 70,000 patients records, and 76 attributes.
- Random Forest, logistic regression, and decision tree have been applied on the dataset to classify cardiovascular disease detection.

Many articles and blogs were perused regarding the application of machine learning to cardiovascular disease detection. Most of them detected cardiovascular disease with dataset and antiquated machine learning techniques. Moreover, the preprocessing part was absent from several publications. Because of this reason, the results were not effective. Consequently, we tried applying data preprocessing with innovation and advanced machine models on the chosen dataset. This research's novelty is using machine learning models like random forest, decision tree, and logistic regression with advanced techniques to construct a system that automatically detects cardiovascular disease using the collected dataset with better results compared to existing works.

3. Proposed Methodology

This section has addressed machine learning models, preprocessing, datasets, and others in brief. Figure 1 shows the schematic diagram of our work.

Figure 1. Schematic diagram of our work.

3.1. Dataset

Kaggle is the source of this dataset [2]. There are 12 features in the Heart Disease dataset. The "target" field is used to anticipate cardiovascular disease based on the patient's cardiac condition. It can have one of two possible integer values: 0 if cardiac disease is not present, and 1 if it is.

3.2. Pre-Processing

Preparing initial data for models is called data preparation. This is where the method of creating a machine-learning model begins. This is the initial step of creating a machine learning model. The changes we make to our dataset before supplying it to the program are referred to as pre-processing. One method for transforming the initial information into an error-free data set is data preparation. Put differently, anytime data is obtained in its original state from several sources, it is not suitable for analysis. In data science, this is a challenging and lengthy aspect. Pre-processing data is an essential requirement for improving machine learning's outcome. Data preprocessing refers to the process of evaluating, screening, manipulating, and storing information so that a machine learning algorithm can understand it and make use of the results. The main purposes of data preparation are to fix issues with information, such as missing numbers, enhance the accuracy of the data, and get the information ready for machine learning [48]. We have applied diverse types of pre-processing techniques in this work.

Improving the Quality of Data: Data preparation is crucial to machine learning because it improves the accuracy of data and provides the foundation for accurate predictions. Cleaning, refining, and deleting mistakes, missing values, and deviations from raw data assures that models and further analysis are based on solid foundations [49]. All techniques were applied here.

Managing Missing Data: In the dataset, there are so many missing values. Many steps can be taken to handle missing values. The information gap is efficiently connected by employing techniques like imputation or reduction. It also helped to get better accuracy. Missing values can be filled out using the mean values. In this dataset, rows were deleted to handle the missing values.

Normalisation and Standardization: This stage minimizes the dominance of certain components over others and combines a range of characteristics and indicators into one common structure, allowing fair comparisons. Applying parameter scaling and min-max scaling is vital for creating a dataset free of noise. In this study, feature scaling and min-max scaling are applied appropriately.

Removing Duplicate Data: All rows, which have duplicate values, were removed from the dataset. The dataset is kept accurate by eliminating redundancies, and further analysis produces reliable and useful insights. Duplicate values can be filled out using the mean values. In this dataset, rows were deleted to handle the Duplicate values. Removing duplicate data, 68,975 records were available.

min_sample_split: Random Forest, a Weak Learner, relies on decision trees to make decisions because algorithmic ensembles are weak learners that are descended

from strong learners. The minimal number of deciding tree events required for a node to split is determined by min_sample_split. We have selected 3 as a min_sample_split.

3.3. Feature Selection

When developing an automatic learning model, any Python-using data analyst needs to understand the value of feature selection. In real-life data science instances, it is rare for each parameter in a dataset to aid in the creation of a model. Repeated factors may decrease a classifier's general precision as well as its capacity to make predictions. Furthermore, the addition of more variables increases the general complexity of the model. Feature selection is a method of determining a subset of features from the original collection of features to shrink the space of features while meeting established standards [50]. In the dataset, there was a target column, which would be predicted. This column was removed from the dataset. Here was another feature named sex, which would help to gain a wrong prediction. Thus, this column was also removed from the dataset. Now, 12 features were selected for our training set.

3.4. Data Splitting

The dataset has been divided into 70:30 ratios. In other words, 30% of the total dataset is the testing set, and 70% for training. While the training data set is used to construct the model, the test data set is utilized for evaluating how well the forecasting algorithm performs.

3.5. Machine learning Algorithms

Three have been used in this project. Random forest, logistics regression, and decision tree have been described below.

1) Random Forest: Random Forest is a method of supervised machine learning employed in programs requiring decision trees, such as regression and classification. Random forests are especially useful for managing complicated and sizable datasets, managing multidimensional feature spaces, and offering insights into the significance of individual features. This method is widely used in a variety of fields, including medical care, banking, and image analysis, because of its capacity to minimize overfitting and maintain excellent predictive accuracy [51]. The random forest method of classification generates a set of decision trees by choosing a random subset of the initial training data. An assortment of decision trees selected randomly from the training set acts as the initial collection. It exceeds the votes from each choice tree to determine the final forecast. A random forest's hyperparameters are like those of a decision tree or bagging classifier. Fortunately, the predictor-class of random forest may be used with ease, eliminating the requirement to mix the decision tree with a sweeping classifier. By using the technique's regressor, random forest may be employed as well to handle regression troubles. Random forest offers more uncertainty to the model as the trees get larger. Rather

than focusing on the most significant characteristic when splitting a node, it requires the most advantageous characteristic from a random collection of characteristics. As a result, the model grows much more varied and is superior. As such, only a randomly selected portion of the characteristics is considered when splitting a node in a random forest classifier. A different approach for increasing the unpredictability of trees is to seek for the best feasible criteria and use arbitrary criteria for each feature [52]. **Figure 2** represents the algorithm of random forest.

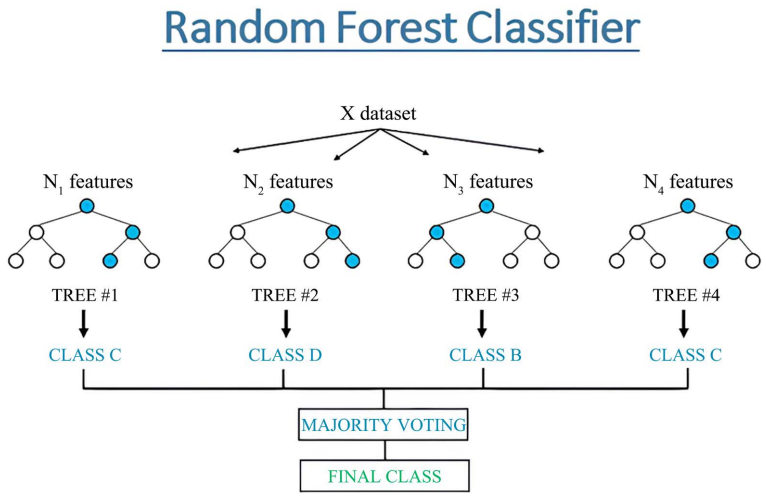


Figure 2. Random forest classifier [53].

2) Logistics Regression: One supervised machine learning technique that is employed for classification problems is known as logistic regression. The aim of this method is to determine the probability that an instance in question belongs to a particular group or not. A procedure used in statistics to examine the connection between two data components is called logistic regression [54]. Because it avoids the limitations with regression methods in this situation and yields results that can be understood as probabilities, logistic regression is especially made to address categorization challenges. The logistic regression's value must, according to the criterion, fall between 0 and 1. On a graph, its restrictions cause it to curve in the shape of a "S" since it can never be greater than 1. Finding the logistic or sigmoid function is made simple with this method.

The threshold value is the concept utilized in relation to Logistic Regression. The chance of either 0 or 1 is defined in part by the threshold values. ones that are beyond the threshold value, for instance, trend to 1, and ones that are below it, to 0 [55]. The premise of linear regression is that both independent and dependent variables have a linear relationship. The line that best fits any number of variables is used. The goal of a linear regression analysis is to forecast the variable that is dependent, which is continuous, with precision. However, the outcome of logistic regression remains between 0 and 1 since it forecasts the likelihood of an event or category that depends upon other variables [55]. The logistic regression algorithm is shown in **Figure 3**.

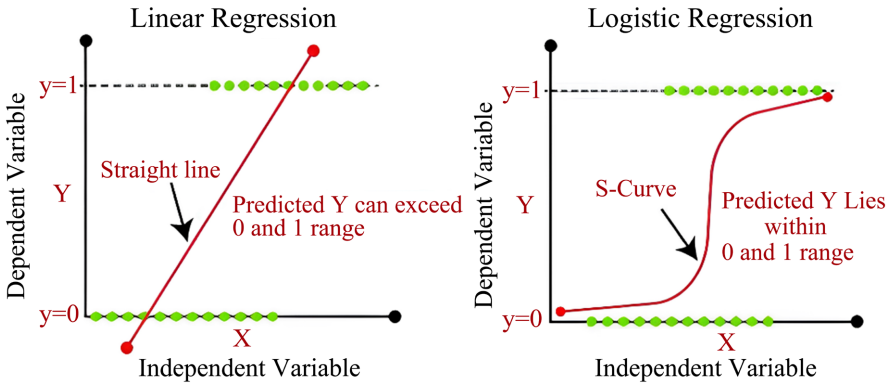


Figure 3. Logistic regression classifier [56].

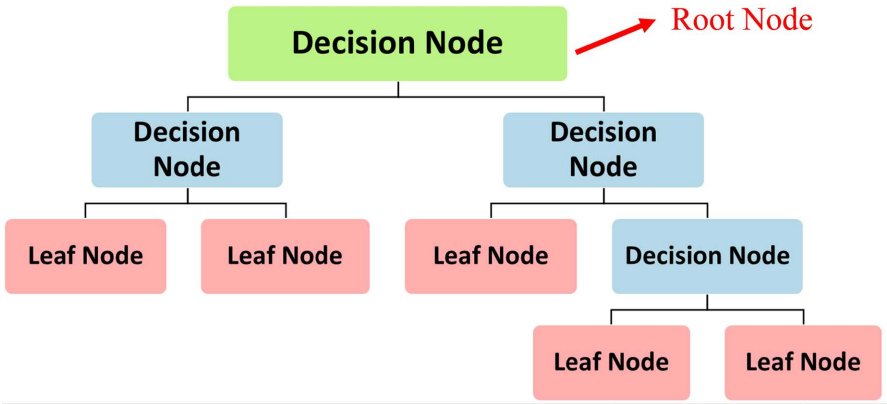


Figure 4. Decision tree classifier [59].

3) Decision Tree: A common and effective tool in many domains, including statistics, data mining, and machine learning, are decision trees. It is especially useful to make a model which can predict the necessary result. Their ability to represent the interactions between numerous factors gives them a clear and understandable means of facilitating data-driven decision-making. The definition, operation, benefits, drawbacks, and uses of decision trees are the main topics of this article. This framework is used for predicting or making decisions, and it appears a lot like a diagram. It includes nodes on the leaf that show the results or forecasts derived from the examinations or choices, branching that shows the result of these experiments or choices, and nodes that reflect characteristic evaluations or choices. Each leaf node indicates a category categorization or an ongoing importance, every inner node reflects a characteristic test, and each branch illustrates the test's result [57]. A decision tree's method begins at the base node and predicts the dataset's class. It is very crucial to make a clear prediction. By matching the contents of the beginning assets with the values of the data (real dataset) characteristic, this method maintains the branch and continues to the following node. The following node in the process arrives by again comparing its characteristic value with those of the remaining sub-nodes. This is what occurs until it arrives at the leaf node of the tree. Decision tree learning employs a divide-and-

conquer strategy through greedy attempts for the most suitable division points inside a tree. After providing a class label to all or most of the records, the division procedure occurs recursively and top-down. The categorization of all the data elements into separate sets or not relies critically on the decision tree's architecture. For smaller trees, it is easier to gather data in a single class, or pure leaf nodes [58]. Figure 4 represents the algorithm of the decision tree.

Model Evaluation: In order to guarantee the resilience and applicability of our approach, we employed many assessment methods. With k = 5, we specifically used k-fold cross-validation, splitting this data set into five subsets. Four subsets per fold were used for training, while one extra subset was used for testing. The rotation of the sample sets among folds reduced the possibility of the overfitting to a particular group of data and enabled us to assess the model's performance thoroughly. We also employed stratified cross-validation at k folds, which keeps the intended class distribution constant throughout each fold, to address possible class imbalances. In heart data sets, where disparities in class may affect the model's prediction accuracy, this step is very important.

4. Results and Discussion

In order to detect cardiovascular disease, machine learning techniques were applied in this investigation. 70,000 patient records were taken from the Kaggle dataset. We used a range of data mining strategies, feature selection, binary classification, data cleaning, and other preprocessing techniques to avoid overfitting. In addition, decision tree, random forest, and logistic regression analysis were performed on the dataset. Ultimately, the accuracy of this model was higher than that of earlier attempts. With the gathered dataset, decision trees, random forests, and logistic regression all functioned effectively. On the training set, a variety of preprocessing techniques were applied, and feature selection helped to increase accuracy. All those techniques combined to provide the exceptional accuracy levels. Table 1 shows the accuracies of three used models. Random forest achieved the highest accuracy, which is 96%.

Table 1. Accuracy table.

Models	Accuracy
Decision Tree	94%
Logistic Regression	80%
Random Forest	**96%**

4.1. Confusion Matrix

A confusion matrix gives a precise picture of how well a categorization model performs in artificial intelligence. All counts are included: false positives, real positives, false negatives, and true negatives. By comparing the expected and true labels, a confusion matrix is a tabular representation that offers a summary of the

performance of a classification model. It displays the proportion of TP, FN, FP, and TN predictions made by the model. This matrix supports model performance analysis by highlighting inaccurate classifications and improving forecast precision. A confusion matrix is a N × N matrix used to evaluate the performance of a classification model, where N is the total number of target classes. The anticipated values of the artificial intelligence model are compared to the goal values in the matrix that really exist. This gives us a thorough grasp of the various kinds of errors and performance indicators related to our classification model.

4.2. Accuracy

It is a discrete metric that cannot be directly optimized. Accuracy is a measure of how frequently a deep learning model predicts the outcome accurately. It can be computed by dividing the total number of guesses by the number of correct forecasts. Accuracy is a metric that can be used to characterize the model's performance in all classes, which is useful when all classes are equally important. More numbers indicate greater model performance.

Figure 5 shows the precision, recall, F1 score, and accuracy results of the decision tree.

	precision	recall	f1-score	support
0	0.95	0.90	0.92	68
1	0.99	0.95	0.97	73
2	0.98	1.00	0.99	64
3	0.82	0.91	0.86	55
accuracy			0.94	260
macro avg	0.94	0.94	0.94	260
weighted avg	0.94	0.94	0.94	260

Figure 5. Accuracy of decision tree.

Figure 6 shows the precision, recall, F1 score, and accuracy results of logistic regression.

	precision	recall	f1-score	support
0	0.90	0.76	0.83	68
1	0.83	0.74	0.78	73
2	0.98	1.00	0.99	64
3	0.53	0.69	0.60	55
accuracy			0.80	260
macro avg	0.81	0.80	0.80	260
weighted avg	0.82	0.80	0.81	260

Figure 6. Accuracy of logistic regression.

Figure 7 shows the precision, recall, F1 score, and accuracy results of random forest classifier.

	precision	recall	f1-score	support
0	0.98	0.94	0.96	68
1	0.97	0.97	0.97	73
2	0.98	1.00	0.99	64
3	0.89	0.93	0.91	55
accuracy			0.96	260
macro avg	0.96	0.96	0.96	260
weighted avg	0.96	0.96	0.96	260

Figure 7. Accuracy of random forest.

Figure 8 represents a clear picture of the accuracies of three models. Decision tree, random forest, and logistic regression performed well. But random forest obtained 96% accuracy, which was the highest accuracy compared to others. It enhanced the prediction power of the system.

Figure 8. Accuracy of three models.

Table 2. Comparison of this work with existing systems.

Author	Dataset	Network	Accuracy
[4]	Healthcare	Decision tree, XGB	0.73
[5]	Kaggle (Heart Disease)	Logistic regression, SVM	0.87
[6]	Kaggle (Heart Disease)	KNN	0.9
[7]	Cleveland Heart Disease database	Neural Network	0.90
[8]	Cleveland Heart Disease database	Decision Tree	0.87
[9]	Kaggle	Random Forest	0.89
[10]	Cleveland Heart Disease database	ANN	0.91
[11]	Own dataset	Naive Bayes	0.78
[12]	Kaggle (Heart Disease)	Random forest	0.92
[13]	Kaggle (Heart Disease)	XGBoost	0.93
[14]	Kaggle (Heart Disease)	Random Forest	0.92
This Work	**Kaggle (Heart Disease)**	**Random Forest**	**0.96**

Table 2 displays the comparison of results obtained for the three types of accuracy and performance across the datasets. Kaggle and Cleveland Heart Disease databases are the popular dataset for detecting cardiovascular disease and most of the papers used it. Kaggle (heart disease) was used to obtain a good outcome and we successfully achieved 96% accuracy using random forest, which one is better than the other. Table 2 represents a good picture of comparison.

5. Conclusions

Finally, we built three models with better accuracy using machine learning, which can easily detect cardiovascular disease with the help of a large dataset. To develop technology for the early detection of cardiovascular disease, the Kaggle dataset was gathered.

To prevent overfitting, a variety of preprocessing techniques were used, including feature selection, binary classification, data cleaning, and data mining. Decision tree, logistic regression, and random forest were applied on the dataset to gain our goal. The random forest model was used to gain a good outcome. In our work, we implemented advanced data preprocessing, data cleaning, and popular classifiers for training our models. Three models were used for training the dataset. Several preprocessing techniques were employed to make our dataset noise free. After training three models on the collected dataset, the outstanding accuracies were achieved. As a result, random forest achieved better accuracy than the other models, which was 0.96. Our models can detect cardiovascular disease quickly within a very low-cost range, revolutionizing medical science.

In the future, we will increase the accuracy of the models used in this work. We will also include or train more models to build a project with better accuracy.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] World Health Organization (n.d.) Cardiovascular Diseases. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[2] Lapp, D. (n.d.) Heart Disease Dataset. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

[3] American Heart Association (2024) More than Half of US Adults Don't Know Heart Disease Is Leading Cause of Death, Despite 100-Year Reign.

[4] Abdelnabi, S., Hasan, R. and Fritz, M. (2022) Open-Domain, Content-Based, Multi-Modal Fact-Checking of Out-of-Context Images via Online Resources. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 14920-14929. https://doi.org/10.1109/cvpr52688.2022.01452

[5] Akter, J., Nilima, S.I., Hasan, R., Tiwari, A., Ullah, M.W. and Kamruzzaman, M. (2024) Artificial Intelligence on the Agro-Industry in the United States of America. *AIMS Agriculture and Food*, **9**, 959-979. https://doi.org/10.3934/agrfood.2024052

[6] Alahmari, T.S., Ashraf, J., Sobuz, M.H.R. and Uddin, M.A. (2024) Predicting the

Compressive Strength of Fiber-Reinforced Self-Consolidating Concrete Using a Hybrid Machine Learning Approach. *Innovative Infrastructure Solutions*, **9**, Article No. 446. https://doi.org/10.1007/s41062-024-01751-8

[7] Amon, M.J., Hasan, R., Hugenberg, K., Bertenthal, B.I. and Kapadia, A. (2020) Influencing Photo Sharing Decisions on Social Media: A Case of Paradoxical Findings. 2020 *IEEE Symposium on Security and Privacy* (*SP*), San Francisco, 18-21 May 2020, 1350-1366. https://doi.org/10.1109/sp40000.2020.00006

[8] Bhuyan, M.K., Kamruzzaman, M., Nilima, S.I., KHATOON, R. and Mohammad, N. (2024) Convolutional Neural Networks Based Detection System for Cyber-Attacks in Industrial Control Systems. *Journal of Computer Science and Technology Studies*, **6**, 86-96. https://doi.org/10.32996/jcsts.2024.6.3.9

[9] Biswas, B., Sharmin, S., Hossain, M.A., Alam, M.Z. and Sarkar, M.I. (2024) Risk Analysis-Based Decision Support System for Designing Cybersecurity of Information Technology. *Journal of Business and Management Studies*, **6**, 13-22. https://doi.org/10.32996/jbms.2024.5.6.3

[10] Emre, G., Akkus, A. and Karamış, M.B. (2018) Wear Resistance of Polymethyl Methacrylate (PMMA) with the Addition of Bone Ash, Hydroxylapatite and Keratin. *IOP Conference Series*: *Materials Science and Engineering*, **295**, Article 012004. https://doi.org/10.1088/1757-899x/295/1/012004

[11] Hasan, R., Al Mahmud, M.A., Farabi, S.F., Akter, J. and Johora, F.T. (2024) Unsheltered: Navigating California's Homelessness Crisis. *Sociology Study*, **14**, 143-156. https://doi.org/10.17265/2159-5526/2024.03.002

[12] Hasan, R., Chy, M.A.R., Johora, F.T., Ullah, M.W. and Saju, M.A.B. (2024) Driving Growth: The Integral Role of Small Businesses in the U.S. Economic Landscape. *American Journal of Industrial and Business Management*, **14**, 852-868. https://doi.org/10.4236/ajibm.2024.146043

[13] Hasan, R., Farabi, S.F., Kamruzzaman, M., Bhuyan, M.K., Nilima, S.I. and Shahana, A. (2024) AI-Driven Strategies for Reducing Deforestation. *The American Journal of Engineering and Technology*, **6**, 6-20. https://doi.org/10.37547/tajet/volume06issue06-02

[14] Hossain, M.A., Tiwari, A., Saha, S., Ghimire, A., Imran, M.A.U. and Khatoon, R. (2024) Applying the Technology Acceptance Model (TAM) in Information Technology System to Evaluate the Adoption of Decision Support System. *Journal of Computer and Communications*, **12**, 242-256. https://doi.org/10.4236/jcc.2024.128015

[15] Johora, F.T., Hasan, R., Farabi, S.F., Akter, J. and Mahmud, M.A.A. (2024) AI-Powered Fraud Detection in Banking: Safeguarding Financial Transactions. *The American Journal of Management and Economics Innovations*, **6**, 8-22. https://doi.org/10.37547/tajmei/volume06issue06-02

[16] Johora, F.T., Manik, M.M.T.G., Tasnim, A.F., Nilima, S.I. and Hasan, R. (2024) Advanced-Data Analytics for Understanding Biochemical Pathway Models. *American Journal of Computing and Engineering*, **4**, 21-34. https://doi.org/10.47672/ajce.2451

[17] Görgün, E. (2022) Characterization of Superalloys by Artificial Neural Network Method. *New Trends in Mathematical Sciences*, **10**, 95-99. https://doi.org/10.20852/ntmsci.2022.470

[18] Akkus, A. (2015) The Investigation of Mechanical Behaviors of Poly Methyl Methacrylate (PMMA) with the Addition of Bone Ash, Hydroxyapatite and Keratin. *Advances in Materials*, **4**, 16-19. https://doi.org/10.11648/j.am.20150401.14

[19] Manik, M.M.T.G., Nilima, S.I., Mahmud, M.A.A., Sharmin, S. and Hasan, R. (2024) Discovering Disease Biomarkers in Metabolomics via Big Data Analytics. *American*

Journal of Statistics and Actuarial Sciences, **4**, 35-49.
https://doi.org/10.47672/ajsas.2452

[20] Al Mahmud, M.A., Hossain, M.A., Saju, M.A.B., Ullah, M.W., Hasan, R. and Suzer, G. (2024) Information Technology for the Next Future World: Adoption of It for Social and Economic Growth: Part II. *International Journal of Innovative Research in Technology*, **10**, 742-747.

[21] Mohammad, N., Khatoon, R., Nilima, S.I., Akter, J., Kamruzzaman, M. and Sozib, H.M. (2024) Ensuring Security and Privacy in the Internet of Things: Challenges and Solutions. *Journal of Computer and Communications*, **12**, 257-277.
https://doi.org/10.4236/jcc.2024.128016

[22] Nilima, S.I., Bhuyan, M.K., Kamruzzaman, M., Akter, J., Hasan, R. and Johora, F.T. (2024) Optimizing Resource Management for IoT Devices in Constrained Environments. *Journal of Computer and Communications*, **12**, 81-98.
https://doi.org/10.4236/jcc.2024.128005

[23] Mohammad, N., Imran, M.A.U., Prabha, M., Sharmin, S. and Khatoon, R. (2024) Combating Banking Fraud with It: Integrating Machine Learning and Data Analytics. *The American Journal of Management and Economics Innovations*, **6**, 39-56.
https://doi.org/10.37547/tajmei/volume06issue07-04

[24] Hasan, R., Farabi, S.F., Al Mahmud, M.A., Akter, J. and Hossain, M.A. (2024) Information Technologies for the Next Future World: Implications, Impacts and Barriers: Part-I. *International Journal of Creative Research Thoughts*, **12**, a323-a330.

[25] Saha, S., Ghimire, A., Manik, M.M.T.G., Tiwari, A. and Imran, M.A.U. (2024) Exploring Benefits, Overcoming Challenges, and Shaping Future Trends of Artificial Intelligence Application in Agricultural Industry. *The American Journal of Agriculture and Biomedical Engineering*, **6**, 11-27.
https://doi.org/10.37547/tajabe/volume06issue07-03

[26] Shahana, A., Hasan, R., Farabi, S.F., Akter, J., Mahmud, M.A.A., Johora, F.T., *et al.* (2024) AI-Driven Cybersecurity: Balancing Advancements and Safeguards. *Journal of Computer Science and Technology Studies*, **6**, 76-85.
https://doi.org/10.32996/jcsts.2024.6.2.9

[27] Sharmin, S., Khatoon, R., Prabha, M., Mahmud, M.A.A. and Manik, M.M.T.G. (2024) A Review of Strategic Driving Decision-Making through Big Data and Business Analytics. *European Journal of Technology*, **7**, 24-37. https://doi.org/10.47672/ejt.2453

[28] Ullah, M.W., Rahman, R., Nilima, S.I., Tasnim, A.F. and Aziz, M.B. (2024) Health Behaviors and Outcomes of Mobile Health Apps and Patient Engagement in the Usa. *Journal of Computer and Communications*, **12**, 78-93.
https://doi.org/10.4236/jcc.2024.1210007

[29] Nikam, A., Bhandari, S., Mhaske, A. and Mantri, S. (2020) Cardiovascular Disease Prediction Using Machine Learning Models. 2020 *IEEE Pune Section International Conference* (*PuneCon*), Pune, 16-18 December 2020, 22-27.
https://doi.org/10.1109/punecon50868.2020.9362367

[30] Johora, F.T., Hasan, R., Farabi, S.F., Alam, M.Z., Sarkar, M.I. and Al Mahmud, M.A. (2024) AI Advances: Enhancing Banking Security with Fraud Detection. 2024 *First International Conference on Technological Innovations and Advance Computing* (*TIACOMP*), Bali, 29-30 June 2024, 289-294.
https://doi.org/10.1109/tiacomp64125.2024.00055

[31] Kamruzzaman, M., Bhuyan, M.K., Hasan, R., Farabi, S.F., Nilima, S.I. and Hossain, M.A. (2024) Exploring the Landscape: A Systematic Review of Artificial Intelligence Techniques in Cybersecurity. 2024 *International Conference on Communications, Computing, Cybersecurity, and Informatics* (*CCCI*), Beijing, 16-18 October 2024, 1-

6. https://doi.org/10.1109/ccci61916.2024.10736474

[32] Ali Linkon, A., Rahman Noman, I., Rashedul Islam, M., Chakra Bortty, J., Kumar Bishnu, K., Islam, A., *et al.* (2024) Evaluation of Feature Transformation and Machine Learning Models on Early Detection of Diabetes Mellitus. *IEEE Access*, **12**, 165425-165440. https://doi.org/10.1109/access.2024.3488743

[33] Habibur Rahman Sobuz, M., Khan, M.H., Kawsarul Islam Kabbo, M., Alhamami, A.H., Aditto, F.S., Saziduzzaman Sajib, M., *et al.* (2024) Assessment of Mechanical Properties with Machine Learning Modeling and Durability, and Microstructural Characteristics of a Biochar-Cement Mortar Composite. *Construction and Building Materials*, **411**, Article 134281. https://doi.org/10.1016/j.conbuildmat.2023.134281

[34] Datta, S.D., Islam, M., Rahman Sobuz, M.H., Ahmed, S. and Kar, M. (2024) Artificial Intelligence and Machine Learning Applications in the Project Lifecycle of the Construction Industry: A Comprehensive Review. *Heliyon*, **10**, e26888. https://doi.org/10.1016/j.heliyon.2024.e26888

[35] Jabin, J.A., Khondoker, M.T.H., Sobuz, M.H.R. and Aditto, F.S. (2024) High-Temperature Effect on the Mechanical Behavior of Recycled Fiber-Reinforced Concrete Containing Volcanic Pumice Powder: An Experimental Assessment Combined with Machine Learning (ML)-Based Prediction. *Construction and Building Materials*, **418**, Article 135362. https://doi.org/10.1016/j.conbuildmat.2024.135362

[36] Khan, M.M.H., Sobuz, M.H.R., Meraz, M.M., Tam, V.W.Y., Hasan, N.M.S. and Shaurdho, N.M.N. (2023) Effect of Various Powder Content on the Properties of Sustainable Self-Compacting Concrete. *Case Studies in Construction Materials*, **19**, e02274. https://doi.org/10.1016/j.cscm.2023.e02274

[37] Sobuz, M.H.R., Joy, L.P., Akid, A.S.M., Aditto, F.S., Jabin, J.A., Hasan, N.M.S., *et al.* (2024) Optimization of Recycled Rubber Self-Compacting Concrete: Experimental Findings and Machine Learning-Based Evaluation. *Heliyon*, **10**, e27793. https://doi.org/10.1016/j.heliyon.2024.e27793

[38] Chinnasamy, P., Arun Kumar, S., Navya, V., Lakshmi Priya, K. and Sruthi Boddu, S. (2022) Machine Learning Based Cardiovascular Disease Prediction. *Materials Today: Proceedings*, **64**, 459-463. https://doi.org/10.1016/j.matpr.2022.04.907

[39] Ali, M.M., Paul, B.K., Ahmed, K., Bui, F.M., Quinn, J.M.W. and Moni, M.A. (2021) Heart Disease Prediction Using Supervised Machine Learning Algorithms: Performance Analysis and Comparison. *Computers in Biology and Medicine*, **136**, Article 104672. https://doi.org/10.1016/j.compbiomed.2021.104672

[40] Dangare, C.S. and S. Apte, S. (2012) Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques. *International Journal of Computer Applications*, **47**, 44-48. https://doi.org/10.5120/7228-0076

[41] Purushottam, Saxena, K. and Sharma, R. (2016) Efficient Heart Disease Prediction System. *Procedia Computer Science*, **85**, 962-969. https://doi.org/10.1016/j.procs.2016.05.288

[42] Rana, M., Ur Rehman, M.Z. and Jain, S. (2022) Comparative Study of Supervised Machine Learning Methods for Prediction of Heart Disease. 2022 *IEEE VLSI Device Circuit and System* (*VLSI DCS*), Kolkata, 26-27 February 2022, 295-299. https://doi.org/10.1109/vlsidcs53788.2022.9811495

[43] Battula, K., Durgadinesh, R., Suryapratap, K. and Vinaykumar, G. (2021) Use of Machine Learning Techniques in the Prediction of Heart Disease. 2021 *International Conference on Electrical*, *Computer*, *Communications and Mechatronics Engineering* (*ICECCME*), Mauritius, 7-8 October 2021, 1-5. https://doi.org/10.1109/iceccme52200.2021.9591026

[44] Ahmed, M.R., Hasan Mahmud, S.M., Hossin, M.A., Jahan, H. and Haider Noori, S.R. (2018) A Cloud Based Four-Tier Architecture for Early Detection of Heart Disease with Machine Learning Algorithms. 2018 *IEEE 4th International Conference on Computer and Communications* (*ICCC*), Chengdu, 7-10 December 2018, 1951-1955. https://doi.org/10.1109/compcomm.2018.8781022

[45] Sobuz, M.H.R., Al-Imran, Datta, S.D., Jabin, J.A., Aditto, F.S., Sadiqul Hasan, N.M., *et al.* (2024) Assessing the Influence of Sugarcane Bagasse Ash for the Production of Eco-Friendly Concrete: Experimental and Machine Learning Approaches. *Case Studies in Construction Materials*, **20**, e02839. https://doi.org/10.1016/j.cscm.2023.e02839

[46] Datta, S.D., Sarkar, M.M., Rakhe, A.S., Aditto, F.S., Sobuz, M.H.R., Shaurdho, N.M.N., *et al.* (2024) Analysis of the Characteristics and Environmental Benefits of Rice Husk Ash as a Supplementary Cementitious Material through Experimental and Machine Learning Approaches. *Innovative Infrastructure Solutions*, **9**, Article No. 121. https://doi.org/10.1007/s41062-024-01423-7

[47] Hasan, N.M.S., Sobuz, M.H.R., Shaurdho, N.M.N., Meraz, M.M., Datta, S.D., Aditto, F.S., *et al.* (2023) Eco-Friendly Concrete Incorporating Palm Oil Fuel Ash: Fresh and Mechanical Properties with Machine Learning Prediction, and Sustainability Assessment. *Heliyon*, **9**, e22296. https://doi.org/10.1016/j.heliyon.2023.e22296

[48] Novogroder, I. (2024) Data Preprocessing in Machine Learning: Steps & Best Practices. https://lakefs.io/blog/data-preprocessing-in-machine-learning/

[49] Goyal, K. (2024) Data Preprocessing in Machine Learning: 7 Easy Steps to Follow. upGrad Blog. https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/

[50] GeeksforGeeks (2024) Feature Selection Techniques in Machine Learning. https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/

[51] GeeksforGeeks (2024) Random Forest Classifier Using Scikit-Learn. https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/

[52] Donges, N., Urwin, M. and Pierre, S. (2024) Random Forest: A Complete Guide for Machine Learning. Built In. https://builtin.com/data-science/random-forest-algorithm

[53] Dinh, A., Miertschin, S., Young, A. and Mohanty, S.D. (2019) A Data-Driven Approach to Predicting Diabetes and Cardiovascular Disease with Machine Learning. *BMC Medical Informatics and Decision Making*, **19**, Article No. 211. https://doi.org/10.1186/s12911-019-0918-5

[54] GeeksforGeeks (2024) Logistic Regression for Classification. https://www.geeksforgeeks.org/understanding-logistic-regression/

[55] Arya, N. (2022) Logistic Regression for Classification. https://www.kdnuggets.com/2022/04/logistic-regression-classification.html

[56] Sreejith, S.S., Rahul, S. and Jisha, R.C. (2016) A Real Time Patient Monitoring System for Heart Disease Prediction Using Random Forest Algorithm. In: *Advances in Signal Processing and Intelligent Recognition Systems*, Springer, Vol. 425, 485-500. https://doi.org/10.1007/978-3-319-28658-7_41

[57] GeeksForGeeks (2024) Decision Tree. https://www.geeksforgeeks.org/decision-tree/

[58] IBM (n.d.) What Is a Decision Tree. https://www.ibm.com/topics/decision-trees

[59] Henglin, M., Stein, G., Hushcha, P.V., Snoek, J., Wiltschko, A.B. and Cheng, S. (2017) Machine Learning Approaches in Cardiovascular Imaging. *Circulation*: *Cardiovascular Imaging*, **10**. https://doi.org/10.1161/CIRCIMAGING.117.005614