

# Comparison of various data mining methods for early diagnosis of human cardiology

Abeer Mohammed Shanshool<sup>1</sup>, Enas Mohammed Hussien Saeed<sup>2</sup>, Hasan Hadi Khaleel<sup>3</sup>

<sup>1</sup>Department of Computer Technologies Engineering, AL-Esraa University Collage, Baghdad, Iraq

<sup>2</sup>Department of Computer Science, College of Education, University of Mustansiriyah, Baghdad, Iraq

<sup>3</sup>Department of Medical Devices Techniques Engineering, AL-Esraa University Collage, Baghdad, Iraq

## Article Info

### Article history:

Received Jul 5, 2022

Revised Sep 12, 2022

Accepted Sep 21, 2022

### Keywords:

Cardiology

Classification algorithms

Cleveland dataset

Data mining

Python

## ABSTRACT

Recent healthcare reports indicate clearly an increasing mortality rates worldwide which puts a significant burden on the healthcare sector due to different diseases. Coronary artery diseases (CAD) is one of the main reasons of these uprising death rates since it affects the heart directly. For early diagnosis and treatment of CADs, a swiftly growing technology called data mining has been used to collect and categorize necessary data from patients; age, blood sugar and pressure, a type of thorax pain, cholesterol, and so on. Therefore, this paper adopted four data mining methods; decision tree (DT), logistic regression (LR), random forest (RF), and Naïve Bayes (NB) to achieve the goal. The paper utilized the Cleveland dataset along with Python programming language to compare among the four data mining methods in terms of precision, accuracy, recall, and area under the curve. The results illustrated that NB method has the best accuracy of 89.47% compared with previous studies which will help with accurate, fast and inexpensive diagnosis of CADs.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Abeer Mohammed Shanshool

Department of Computer Technologies Engineering, AL-Esraa University Collage

Al-Andalus Square, Baghdad, Iraq

Email: Neqolet@yahoo.com

## 1. INTRODUCTION

Arrhythmia and abnormal heartbeats are potential symptoms of acute heart disease and it causes the largest number of deaths worldwide by 2030. The mortality rate will increase to 23.6 million people [1], [2]. This mainly affects men because of their smoking habits. The considered disease types are heart cancer, myocarditis, congenital heart disease, heart failure, coronary artery diseases, cardiomyopathy, angina pectoris, arrhythmia, heart attacks, and so on [3]. Among the risk operators for Cardiology are smoking, age, high blood pressure, family history, malnutrition, and cholesterol [4]. Due to a lack of knowledge, people are unable to detect the heart diseases earlier which leads to many death cases. Mostly, the case is exposed at a late stage or after death since people are unwilling to provide early appropriate treatments due to the high cost of these treatments [5]. When it comes to early prediction of heart diseases, a well-used application utilizes the machine learning technology to achieve the goal. However, the procedure of converting worthless data to helpful is referred to as the data mining [6]. Data mining is a natural development of information technology, especially today with big data being used in discovery and analysis to establish logical relationships; data mining methods are being used in different areas, particularly in the medical field, as performing these methods helps clinicians select treatments, predict patient outcomes, and reduce costs in medicine, and instruct researchers to develop new treatments and refer patients to participate clinical trials [7]. Diagnosis in the medical field plays an important role in saving life, however, it is a complex task to be

performed efficiently and accurately to decrease costs of the clinical testing procedures and suitable computing data support [8]. Significant studies have adopted various classification techniques throughout the history of heart diseases for early prediction and diagnosis of heart diseases.

The purpose of the present study is:

- Solve the classification problem using data mining methods to get the most accurate results in terms of accuracy, recall, sensitivity and a receiver operating characteristic (ROC) to predict heart disease early
- Using four algorithms of machine learning (LR, Naïve Bayesian, DT and RF) to predict cardiology
- Finding best a method for prediction on the cardiology dataset
- Compared with the results of previous studies to prove the effectiveness of the accuracy of our proposed system.

Mary and Sebastian [9] analyzed three sets of data, the aim of which is to consider characteristics that may affect the core, data analysis was performed using several classification methods, including the Naïve Bayesian classifier, random forest (RF) and random tree, where the Weka environment was used for data analysis; results showed an increase in prediction accuracy, especially with the Naïve Bayes and RF algorithms. Suresh *et al.* [10] The aim of this study has been to use a hybrid model combining the support vector machine and RF to predict heart disease, in which repetitive features were removed from the RF to identify features that improve the results of the vector support machine for heart disease prediction, as it was applied to a data set (Pima Indiana heart disease). The two algorithms give a more efficient and effective result with an accuracy of up to 98.3%, unlike if the two algorithms are used separately. In addition, an analysis was performed to consider the effect of the gamma coefficient and the regularization coefficient of the svm algorithm where svm was very sensitive. Sureja *et al.* [11] proposed a new method that combined the slap swarm algorithm with a support vector machine to apply it to two cardiology datasets, first from kaggle and University of California Irvine (UCI) and second from clinical-heart failure records; initially, the slap swarm algorithm was used to choose the best features that could affect the prediction accuracy. Then, the support vector machine (SVMs) algorithm has been used in order to predict heart disease; where performance had been assessed with the use of the accuracy, sensitivity, recall, f-scale and G-mean; The proposed system seemed effective and quick to predict, as it reached an accuracy of 98.75% and 98.46% for two data sets. In addition Afrin *et al.* [12] and Assegie *et al.* [13] conducted study to proposed system to predict liver disease that used machine learning algorithm like (SVM, RF, k-neighbor nearest (KNN), DT and Naïve Bayes (NB)) the experiment result shows DT has acheive accueate by 94.295%; Jasim *et al.* in [14] presented a predictive system for the prognosis of breast cancer using the UCI repository dataset. They used many algorithms like (RF, perceptron, SVM, DT, KNN, NB, multi-layer perceptron (MLP), LR and xgboost); The simulation result showed that the SVM and Perceptron algorithms had a high accuracy of up to 90% compared to the other algorithms used.

Reddy *et al.* in [15] used dataset from UCI repository to predicte breast cancer by using many algorithms such as Adaboost, multi-layer perceptron and Stacking classifier, a stacking algorithm has been successful in achieving the best accuracy 99.20%; Ramanath *et al.* [16] they suggested a hybrid system based on fuzzy logic approach to supporting breast cancer prediction illness; the suggested system had shown a 96.49% classification accuracy and Wisconsin data was used. Blockchain has been used to provide security for ensuring the fact that only trusted and reputable agents participate participation in the process of the decision-making.

Ghosh *et al.* [17] used dataset Wisconsin breast cancer database (WBCD) to introduced a novel approach by using severial algorithm like LR, DT, RF, SVM and KNN and used Least Absolute Shrinkage and Selection Operator (LASSO) as feature selction and result show RF was best accurate by 99.41%; Jittawiriyankoon and Srisarkun [18]. They used location data, machines, statistics, and downtime from a data-mining plant using artificial intelligence and machine learning to develop a decision support strategy; scheduling a maintenance plan by using open source software for replacing the shortcut of maintenance planning and scheduling; on data mining, 3 promising training algorithms are used for insightful data as a result precision numbers have been obtained; Alalwan *et al.* [19] two data mining algorithms were used a self-organizing map and a RF for diabetes diagnosis; the results have shown that they can provide services in the health care sector to make effective decisions; Saranya and Pravin [20] in this work, they comprehensively study the different strategies used to predict diseases by applying several mining algorithms to improve prediction to reduce hospital admissions and reduce patient costs. In this paper, we initially i) allocate four classification methods and choose the most accurate method in predicting and detecting heart disease through data mining techniques, ii) we compare the results of the proposed work with previous studies that used the same data set. The paper had been divided to the following sections; The “Methodology” presents a flowchart of the proposed work with the data set had used and the pre-processing of the data and all details, the “Result and Discussion” of the experimental result and validation, and finally the “Conclusions” the paper concludes.

**2. METHODOLOGY**

**2.1. Flowchart**

The flowchart contains three main steps and as illustrated in Figure 1. It begins with the preparation step where the information is filtered and categorized prior of any processing. The output from stage one goes to stage two which is divided to training and testing sub-stages of the four classification techniques. The last step is where these techniques are evaluated throughout several performance evaluations to eliminate low performances. The main steps for each stage are described in details next.

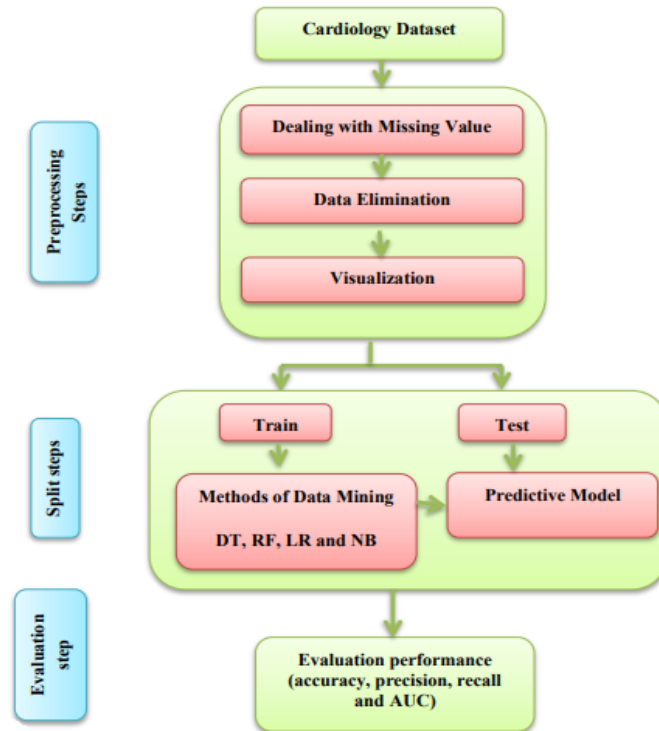


Figure 1. Proposed method flowchart

**2.2. Dataset**

This paper adopted a dataset of real cardiac values which was taken from the Cleveland dataset [21]. This dataset provides information on risk factors of heart diseases. It contains 303 cases and 14 attributes. Patients are classified either infected or uninfected. The number of people without heart issues are 138, and those with heart conditions are 165 samples, as shown in Table 1 and Table 2 illustrates some samples of the dataset.

Table 1. Cardiac dataset description

Attribute	Number of attributes
Total cases	303
Dimensions	13
Category	2
cases per category	infected =165 uninfected = 138
Features	Real

Table 2. Original specimens from the data-set

Age	Gender	cp	Trest bps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
62	0	0	140	268	0	0	160	0	3.6	0	2	2	0
63	1	0	130	254	0	0	147	0	1.4	1	1	3	0

Where:

Age: Refers to the age of patients.

Sex: gender (1 pointing to males; 0 pointing to females).

cp indicates chest pain.

Trestbps signalize to comforting the pressure of the blood.

Chol: indicates the serum cholesterol in (mg/dl).

fbs: fasting blood sugar higher than 120mg /dl (0 is False; 1 is True).

Restecg: pointing to restecg which signalizes the resting electro cardio graphic results.

Thalach: Highest heart rate.

Exang: indicates the exercise induced angina (1 = yes) and (0 = no)

Oldpeak: ST depression that has been induced by the exercise relative to rest

Slope: refers to slope of peak exercise ST segment

ca: major number for vessels that have been colored by the fluoroscopy (0 - 3)

Thal: (3 indicates normal; 6 represents fixed defect and 7 indicates reversible defected).

Target: pointing to heart disease diagnosis (1 = infected; 0 = uninfected)

### 2.3. Data pre-processing [22]

It is of high importance in the data mining method by which the raw data is converted into a recognizable format. Often, the real-world data is inconsistent, incomplete, lacks certain trends or behaviors, and possibly Contains various errors. So, this step is important to solve this type of problems because it helps with detecting anomalies. The quality of used data should be of high accuracy in order to be located within the dataset. The data should also be with no noise or errors because these issues can affect directly the model’s ability of learning. Throughout preprocessing, the data goes through series of steps: Handling missing values: The data is cleaned up via processes like deleting rows with missing data or filling missing values. Data estimation: refers to the verification of any missing value or noise in the data because, in most interpretations, it cannot be done with missing data. Data visualization: helps to clarify the relationship between features and their interpretation, Figure 2 illustrates a visualization of features. After the pre-processing step, the dataset is broken into 2 parts; 75% for training and 25% for testing. Training process is for the four classification methods; LR, DT, RF and NB (they used to build the classification model) and the testing process is set to test the prediction model.

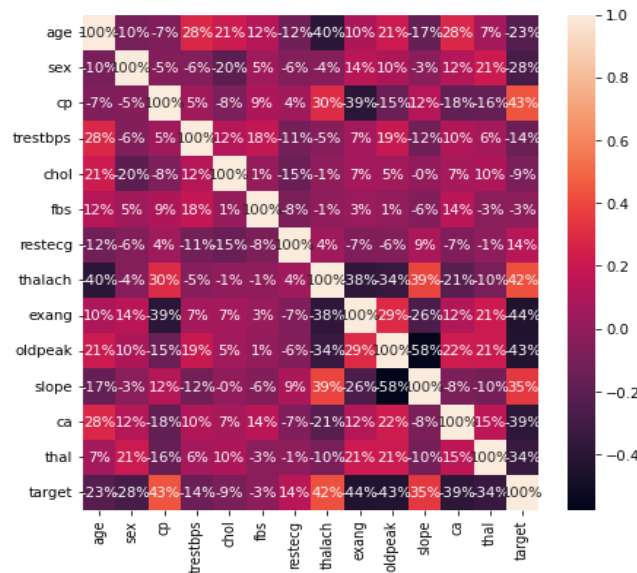


Figure 2. Visualization between attributes

### 2.4. Performance evaluation

After completing the training and testing phases of the four classification methods, the process continues to the final step which is performance appraisal. Where a confusion matrix [23] is used as a useful evaluation method that contains many values to measure performance; like accuracy, area under curve of ROC or area under curve (AUC), recall and precision [24]–[28], which are intended for the validation step.

Without validation, we only have information about how the models perform with the training data [29]. The area under ROC curve is referred to as AUC and is often used for erroneous ratings. The value ranges from 0 to 0 to 1.0, while the real value of 1.0 of the equivalents represents the ideal score and the value 0.5 represents a random prediction [30].

$$\text{Accuracy} = \frac{x+d}{x+y+c+d} \tag{1}$$

$$\text{Precision (p)} = \frac{x}{x+c} \tag{2}$$

$$\text{Recall (r)} = \frac{x}{x+y} \tag{3}$$

Where:

- *x* indicates to true positive (TP): represents the number of correctly labeled positive examples.
- *y* indicates to false negative (FN): represents number of incorrectly classified negative samples.
- *c* indicates to false positive (FP): represents number of incorrectly classified positive samples.
- *d* indicates to true negative (TN): represents number of negative samples which were correctly classified.

### 3. RESULTS AND DISCUSSION

We applied four classification algorithms after removing the anomalies from the heart disease data to compute the performance parameters for the proposed algorithms. The classification algorithms that applied with the proposed work are RF, NB, LR, and DT to evaluate algorithms for accuracy, precision, and recall and under area curve. Table 3 lists results of performance comparison among the four classification algorithms.

Table 3. A comparison to evaluate the performance of algorithms

Algorithms	Accuracy	Precision	Recall	AUC
Logistic regression	85.53%	86.36%	88.37%	92.60%
Decision tree	81.58%	80.85%	88.37%	77.58%
Naïve Bayes	89.47%	88.88%	93.02%	<b>93.44%</b>
Random forest	86.84%	90.24%	86.04%	92.60%

Table 3 clearly demonstrated that NB has the optimal performance results in the terms of accuracy, recall, and AUC. In the meanwhile, RF scored better in terms of precision by 90.24 which improves the effectiveness of accuracy upon removing errors and anomalies from the cardiac data. The results of Table 4 are represented in a chart graph and as shown in Figure 3. From the graph it may be seen the NB has highest accuracy by 89.47%. It is worth to mention that NB is a method of always high accuracy but this is not the case with other methods since they depend on specific features.

Table 4. Compared previous work with proposed algorithms

Reference	Algorithms	Accuracy	Precision	Recall	ROC
Ekiz <i>et al.</i> (2017) [1]	DT	67.6%	-	-	-
	RF	77.42%	-	-	-
Kohli <i>et al.</i> (2018) [31]	DT	70.97%	-	-	-
	DT	76.66%	-	-	-
Maji <i>et al.</i> (2019) [32]	NB	89.2%	-	-	-
Tarawneh <i>et al.</i> (2019) [8]	DT	75.82 %	-	-	-
	RF	76.92%	-	-	-
	LR	80.21%	-	-	-
Pawar <i>et al.</i> (2020) [33]	NB	76.92%	-	-	-
	RF	80.89%	-	-	89.53
	RF	88%	0.87	0.87	91.7
Ripan <i>et al.</i> (2021) [35]	LR	85%	0.86	0.84	92.5
	NB	84%	0.85	0.82	90.0
	DT	81.58%	80.36	88.37	77.58
Proposed algorithms	RF	86.84%	90.24	86.04	92.60
	LR	85.53%	86.36	88.37	92.60
	NB	89.47%	88.88	93.02	93.44

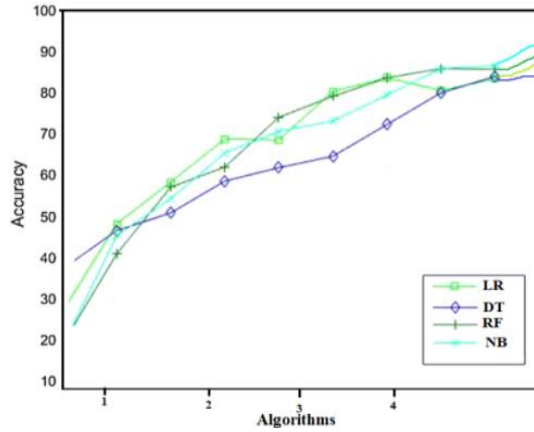


Figure 3. Comparison of accuracy

Figure 4 illustrated a comparison of area under curve among the four methods. It can be seen Figure 4(a) indicate to DT has the lowest (0.7758). NB has the highest AUC (0.9260) as shown in Figure 4(b), and in Figures 4(c) and 4(d) indicate to receiver operating characteristics for RF and LR are equal values (0.9260).

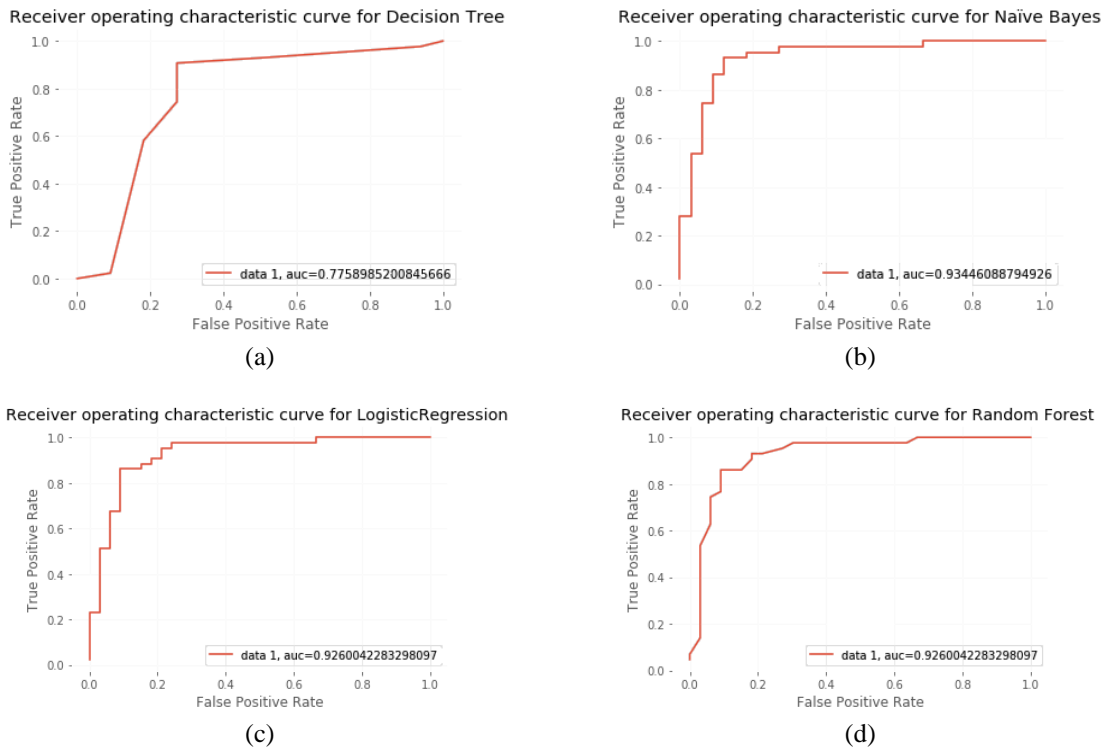


Figure 4. Comparison of area under curve (a) ROC curve for DT, (b) ROC curve for Naive Bayes, (c) ROC curve for LR, and (d) ROC curve for RF

**3.1. Age, sex and target with heart diseases**

Figure 5 illustrated the connection between patients’ ages and heart diseases. It is clearly shown that cardiac issues start to increase in the from 40 to 60 whether a high risk of cardiovascular diseases starts from 60 and above. This is because aging causes alteration in the blood and heart utensils. Thus, increasing the risk of getting a heart issue. Cardiovascular disease is more common in men than women and it remains the leading death cause in men from smoking; Figure 6 shows the prevalence of heart disease between genders

(males and females), with 1 representing males and 0 representing females. Heart failure means the heart is not pumping as well as it should; according to the dataset used, it is classified into two parts: where the number (1) represents the infected cases and (0) represents the uninfected cases; Figure 7 shows the target (0) as it represents 138 uninfected cases and the target (1) is 165 infected cases.

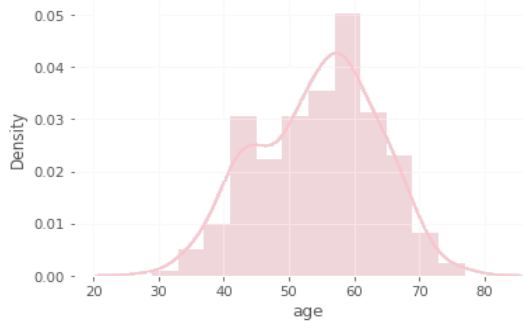


Figure 5. The relationship between age and heart disease

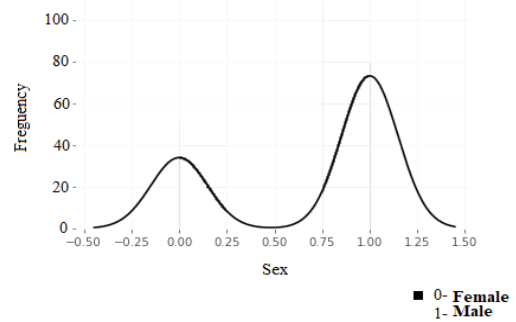


Figure 6. Sex Vs heart disease

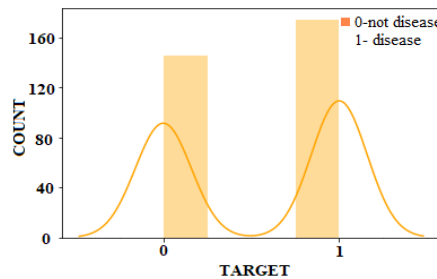


Figure 7. Penitent and heart disease

### 3.2. Validating the suggested approach

The goal of the performance appraisal process is to improve the way the model works to achieve a higher level of efficiency; to validate the proposed procedure. A comparison was made between the proposed work and the most recent published work from the last five years that used the same data set, we used performance metrics as (accuracy, precision, recall and ROC) selected for comparison. The proposed approach together with previous works were tested Cleveland dataset.

The proposed algorithm had evaluated AUC, precision, accuracy, and recall that had been presented as a standard measure in risk measurements by healthcare field; however, higher (AUC) confirms the accuracy [36]. Table 4 shows how the proposed approach performs better than the previous studies. It can predict accuracy with DT by 81.58%, precision by 80.36%; recall by 88.37% and AUC by 77.58% which is better than all the seven methods Listed in the table. Whereas RF achieved 86.84% accuracy, 92.60% AUC, 90.24% precision and 86.04% recall. Moreover, NB achieved 89.47% accuracy; 88.88% precision, 93.02% recall and 93.44% AUC. As for the LR method 85.53%, 86.36, 88.37 and 92.60 for the precision, accuracy, recall and AUC respectively.

### 4. CONCLUSION

In this paper, we applied the most common algorithms in data mining, which is the (DT, RF, NB and LR) algorithms by applying it to the heart data taken from the Cleveland website and available on the Internet, where the data had been trained to 70% training and 30% testing, and then many of scales that have been used in the papers were used. To compare the results with it, such as accuracy, precision, sensitivity, and ROC, after comparing the results with similar work for the last five years; The goal of the proposed system is to use an efficient method of data mining to achieve effective analyzes of medical information that helps in rapid disease prediction for patient care. The results of the work clearly showed that NB has better prediction results among all with a 89.47%. The new method could reduce the number of patients tested, reduce deaths, improve economics, and expand community coverage.

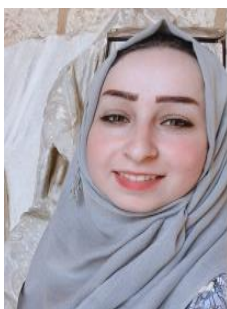
## REFERENCES




- [1] S. Ekiz and P. Erdogmus, "Comparative study of heart disease classification," *2017 Electric Electronics, Computer Science, Biomedical Engineering's Meeting, EBBT 2017*, 2017, doi: 10.1109/EBBT.2017.7956761.
- [2] I. Tougui, A. Jilbab, and J. El Mhamdi, "Heart disease classification using data mining tools and machine learning techniques," *Health and Technology*, vol. 10, no. 5, pp. 1137–1144, 2020, doi: 10.1007/s12553-020-00438-1.
- [3] J. Abdollahi, B. Nouri-Moghaddam, and M. Ghazanfari, "Deep neural network based ensemble learning algorithms for the healthcare system (diagnosis of chronic diseases)," 2021.
- [4] M. N. Kumar, K. V. S. Koushik, and K. Deepak, "Prediction of heart diseases using data mining and machine learning algorithms and tools," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 3, no. 3, pp. 887–898, 2018, doi: 10.13140/RG.2.2.28488.83203.
- [5] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*. Elsevier, 2012.
- [6] D. Jain and V. Singh, "A two-phase hybrid approach using feature selection and adaptive SVM for chronic disease classification," *International Journal of Computers and Applications*, vol. 43, no. 6, pp. 524–536, 2021, doi: 10.1080/1206212X.2019.1577534.
- [7] I. Preethi and K. Dharmarajan, "Diagnosis of chronic disease in a predictive model using machine learning algorithm," *Proceedings of the International Conference on Smart Technologies in Computing, Electrical and Electronics, ICSTCEE 2020*, pp. 191–196, 2020, doi: 10.1109/ICSTCEE49637.2020.9276957.
- [8] M. Tarawneh and O. Embarak, "Hybrid approach for heart disease prediction using data mining techniques," *Lecture Notes on Data Engineering and Communications Technologies*, vol. 29, pp. 447–454, 2019, doi: 10.1007/978-3-030-12839-5\_41.
- [9] T. R. S. Mary and S. Sebastian, "Predicting heart ailment in patients with varying number of features using data mining techniques," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 4, pp. 2675–2681, 2019, doi: 10.11591/ijece.v9i4.pp2675-2681.
- [10] T. Suresh, T. A. Assegie, S. Rajkumar, and N. K. Kumar, "A hybrid approach to medical decision-making: diagnosis of heart disease with machine-learning model," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 2, pp. 1831–1838, 2022, doi: 10.11591/ijece.v12i2.pp1831-1838.
- [11] N. Sureja, B. Chawda, and A. Vasant, "A novel salp swarm clustering algorithm for prediction of the heart diseases," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 1, pp. 265–272, 2022, doi: 10.11591/ijeecs.v25.i1.pp265-272.
- [12] S. Afrin *et al.*, "Supervised machine learning based liver disease prediction approach with LASSO feature selection," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, pp. 3369–3376, 2021, doi: 10.11591/eei.v10i6.3242.
- [13] T. A. Assegie, R. Subhashni, N. K. Kumar, J. P. Manivannan, P. Duraisamy, and M. F. Engidaye, "Random forest and support vector machine-based hybrid liver disease detection," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 3, pp. 1650–1656, 2022, doi: 10.11591/eei.v11i3.3787.
- [14] A. A. Jasim, A. A. Jalal, N. M. Abdulateef, and N. A. Talib, "Effectiveness evaluation of machine learning algorithms for breast cancer prediction," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 3, pp. 1516–1525, 2022, doi: 10.11591/eei.v11i3.3621.
- [15] S. S. Reddy, N. Pilli, P. Voosala, and S. R. Chigurupati, "A comparative study to predict breast cancer using machine learning techniques," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 27, no. 1, pp. 171–180, Jul. 2022, doi: 10.11591/ijeecs.v27.i1.pp171-180.
- [16] T. T. Ramanath, M. J. Hossen, and M. S. Sayeed, "Blockchain integrated multi-agent system for breast cancer diagnosis," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 2, pp. 998–1008, May 2022, doi: 10.11591/ijeecs.v26.i2.pp998-1008.
- [17] P. Ghosh, A. Karim, S. T. Atik, S. Afrin, and M. Saifuzzaman, "Expert cancer model using supervised algorithms with a LASSO selection approach," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 3, pp. 2631–2639, Jun. 2021, doi: 10.11591/ijece.v11i3.pp2631-2639.
- [18] C. Jittawiriyankoon and V. Srisarkun, "Simulation for predictive maintenance using weighted training algorithms in machine learning," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 3, pp. 2839–2846, Jun. 2022, doi: 10.11591/ijece.v12i3.pp2839-2846.
- [19] S. A. D. Alalwan, "Diabetic analytics: Proposed conceptual data mining approaches in type 2 diabetes dataset," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 14, no. 1, pp. 88–89, Apr. 2019, doi: 10.11591/ijeecs.v14.i1.pp88-95.
- [20] G. Saranya and A. Pravin, "A comprehensive study on disease risk predictions in machine learning," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 4, pp. 4217–4225, 2020, doi: 10.11591/ijece.v10i4.pp4217-4225.
- [21] "Heart Disease Dataset," 2022. <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>.
- [22] J. S. Krishnan and S. Geetha, "Prediction of heart disease using machine learning algorithms," in *Proceedings of 1st International Conference on Innovations in Information and Communication Technology, ICICT 2019*, 2019, pp. 1–5, doi: 10.1109/ICICT1.2019.8741465.
- [23] M. Z. Alam, M. S. Rahman, and M. S. Rahman, "A random forest based predictor for medical data classification using feature ranking," *Informatics in Medicine Unlocked*, vol. 15, 2019, doi: 10.1016/j.imu.2019.100180.
- [24] A. Kishor and C. Chakraborty, "Artificial intelligence and internet of things based healthcare 4.0 monitoring system," *Wireless Personal Communications*, 2021, doi: 10.1007/s11277-021-08708-5.
- [25] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.
- [26] D. Deepika and N. Balaji, "Effective heart disease prediction using novel MLP-EBMDA approach," *Biomedical Signal Processing and Control*, vol. 72, 2022, doi: 10.1016/j.bspc.2021.103318.
- [27] A. M. Shanshool, A. H. Salman, A. G. H. Rafash, and E. M. H. Saeed, "A review study on machine learning approaches on coronavirus big data," *Iraqi Academic Scientific Journals*, no. 37, pp. 431–459, 2022.
- [28] I. M. El-Hasnony, O. M. Elzeki, A. Alshehri, and H. Salem, "Multi-label active learning-based machine learning model for heart disease prediction," *Sensors*, vol. 22, no. 3, 2022, doi: 10.3390/s22031184.
- [29] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, R. Sun, and I. García-Magarinõ, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2018, 2018, doi: 10.1155/2018/3860146.
- [30] L. Riyaz, M. A. Butt, M. Zaman, and O. Ayob, "Heart disease prediction using machine learning techniques: a quantitative review," in *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021*, 2022, vol. 3, pp. 81–94, doi: 10.1007/978-981-16-3071-2\_8.






- [31] P. S. Kohli and S. Arora, "Application of machine learning in disease prediction," *2018 4th International Conference on Computing Communication and Automation, ICCCA 2018*, 2018, doi: 10.1109/CCAA.2018.8777449.
- [32] S. Maji and S. Arora, "Decision tree algorithms for prediction of heart disease," *Lecture Notes in Networks and Systems*, vol. 40, pp. 447–454, 2019, doi: 10.1007/978-981-13-0586-3\_45.
- [33] R. Pawar, S., Nanaware, P., Shejwal, S., & Goudar, "Prediction of heart disease using hybrid approach and ensemble learning," *Journal of Physics: Conference Series*, vol. 40, no. 74, pp. 2142–2147, 2020, doi: 10.1088/1742-6596/1916/1/012445.
- [34] S. J. Pasha and E. S. Mohamed, "Novel feature reduction (NFR) model with machine learning and data mining algorithms for effective disease risk prediction," *IEEE Access*, vol. 8, pp. 184087–184108, 2020, doi: 10.1109/ACCESS.2020.3028714.
- [35] R. C. Ripan *et al.*, "A data-driven heart disease prediction model through k-means clustering-based anomaly detection," *SN Computer Science*, vol. 2, no. 2, 2021, doi: 10.1007/s42979-021-00518-7.
- [36] C. Pan, A. Poddar, R. Mukherjee, and A. K. Ray, "Impact of categorical and numerical features in ensemble machine learning frameworks for heart disease prediction," *Biomedical Signal Processing and Control*, vol. 76, 2022, doi: 10.1016/j.bspc.2022.103666.

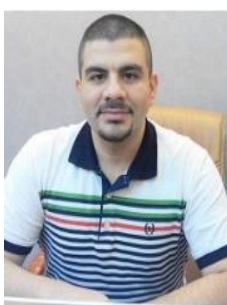
## BIOGRAPHIES OF AUTHORS






**Abeer Mohammed Shanshool**    was born in 1989 in Baghdad. She holds a Bachelor's degree in Computer Science from Al-Mustansiriya University, Baghdad, Iraq, and a Master's degree in Information Technology from Altinbas University, Istanbul, Turkey. Her current research interests lie in the areas of machine learning, big data, fog computing, artificial intelligence, data mining and deep learning. She is working as Assistant Lecture at Al- Esraa University Collage, Baghdad since 2012. She can be contacted at email: Neqolet@yahoo.com.



**Enas Mohamed Hussein Saeed**    was born in 1976 in Iraq. She holds a Bachelor's degree and a Master's degree in Computer Science from Al-Mustansiriya University, Baghdad, Iraq, and PhD from the University of Technology. Her current research interests lie in the areas of machine learning, artificial intelligence and data mining. She is working as Assistant Prof. at Al-Mustansiriya University, Baghdad since 2004. She can be contacted at email: drenasmohammed@uomustansiriyah.edu.iq.



**Hasan Hadi Khaleel**    was born in Iraq 1979. I received my B.Sc. in Electrical Eng. & M.Sc. in Control and Computer Eng. from University of Baghdad, in 2001 and 2003 respectively. I received my PhD in Computer Graphics/Biomedical Engineering from the faculty of Computer Science and Information Technology at University Putra Malaysia/Malaysia in 2012. Among my research interests are biomedical imaging & engineering, computer graphics, artificial intelligence, machine learning, computer assisted surgery, image processing, human-computer interaction, augmented reality, virtual reality, big data, healthcare & technology. He can be contacted at email: hasan.khaleel@esraa.edu.iq.