ISSN: 2832-7705

Research Article

International Journal of Clinical and Medical Education Research

Classification and Diagnosis of Heart Disease Using Machine Learning

Ayedh Abdulaziz Mohsen[1]*, Kharroubi Naoufel[2], Taher Alrashahy[3] and Somia Noaman[4]

[1]*Department of CS and IT, Ibb University, Computer Department, Aljazeera University, Yemen*

[2]*Department of Computer Science, University College of Khurma Taif University, KSA*

[3]*Department of CS Faculty of Computer Science, Hodiadah University, Hodiadah, Yemen*

[4]*Department of CS and IT Faculty of Science, Ibb University, Ibb, Yemen*

***Corresponding Author**
Ayedh Abdulaziz Mohsen, Department of CS and IT, Ibb University, Computer Department, Aljazeera University, Yemen.

Submitted: 2024 , Jun 06; **Accepted**: 2024, Aug 29; **Published**: 2024, Sep 02

Citation: Mohsen, A, A., Naoufel, K., Alrashahy, T., Noaman, S. (2024). Classification and Diagnosis of Heart Disease Using Machine Learning. *Int J Clin Med Edu Res,* 3(5), 01-12.

Abstract

The study aimed to explore the application of machine learning techniques in diagnosing and classifying various types of heart diseases. A number of algorithms commonly used in healthcare, such as the naive Bayes model, SVM, k-nearest neighbour (K-NN), and others, were reviewed. This study highlights the importance of the quality of the data used in the database to obtain an accurate and reliable diagnosis. The data were collected from patient records in hospitals and clinics and were analysed and compared with those of previous relevant studies. Clinical decision assistance software has been used to help surgeons make medical decisions based on patient information. Positive results have been achieved that confirm the effectiveness of using machine learning techniques in diagnosing heart disease. These technologies have shown the potential to improve the accuracy and efficiency of diagnosis, leading to improved patient outcomes and reduced health burdens. The findings also revealed the need to develop effective diagnostic tools and enhance the prevention of heart disease. This study provides an important foundation for healthcare professionals and doctors working in the field of cardiology, as the techniques used can help them better understand and diagnose conditions and improve patient care.

Keywords: Heart Diseases, Classification, Diagnosis, Techniques, Machine Learning

1. Introduction

Currently, heart disease is among the most prevalent diseases worldwide. It has been estimated that it caused approximately 17.9 million people to die in 2017—15% of all deaths that occur naturally [1]. Heart disease is considered a chronic disease that can be detected earlier by measuring various health standards, e.g., glucose level, heart rate, cholesterol, and blood pressure [2]. Heart disease does not affect human health alone; it affects the capabilities of countries and their economies [3]. That is, heart disease is a serious disease with an extremely high incidence rate, particularly in poorer nations, due to the lack of knowledge of its symptoms [4]. Currently, many data mining and deep learning algorithms have been developed to identify and predict various types of diseases [5]. However, there are classification techniques that are widely used in healthcare because they are able to process very large amounts of data [6]. The common techniques used in healthcare are naive bayes, support vector machine (SVM), the k-nearest neighbor algorithm (k-NN), decision tree, fuzzy logic, artificial neural network (ANN), and genetic algorithms (GA) [7]. Systems for heart disease diagnosis and their applications possess both sensitivity and accuracy; however, the latter is dependent on how accurately the data are kept in the database [8]. Clinical decision aids help clinicians make medical decisions based on information provided by patients regarding their symptoms [9]. That is, these programs can make decisions using several features to analyse the input data and/or to arrive at the ultimate output (such as apps that use symptoms to identify diseases). Because many doctors lack enough information about the features of diseases, computer systems, and interconnected technologies that help patients identify symptoms of diseases, in cases where patients have many illnesses, they might not be able to accurately diagnose conditions quickly [10]. However, there are inspection techniques

(e.g., automatic learning, decision trees, neural networks, etc.) that help diagnose and predict disease by using a web application that simulates a neural network algorithm [11].

There are special programs that can be used to train medical students and doctors on new technologies in any field, educate patients, and identify disease symptoms [12]. That is, the system can simulate the patient's symptoms through graphic web applications to diagnose his or her disease. On this basis, this study attempts to compare studies related to heart disease and related features to gather data on the analysis and classification of heart diseases collected from patient documentation in hospitals in Ibb city, Yemen [10]. Therefore, the overall objective of this paper is to design a system to diagnose and classify heart diseases. This, in turn, may lead society to be aware of disease risks, creating a medical culture to be aware of preventive techniques to avoid heart diseases, which include eight kinds of heart diseases, such as artery occlusion, heart-related rheumatism, angina pectoris heart disease, heart-related disorders, birth defects, heart arrhythmia and cardiomyopathy. The paper's structure is arranged as follows: A review of pertinent research in this field is presented in Section II. The materials, procedures, and methodology employed are covered in Section III. The results and conclusions are given in Section IV. In Section V, the study's results are presented, along with some recommendations for further research.

2. Related Works

Before discussing studies that diagnose and classify heart disease, it is important to define heart disease, in which the term “heart disease” refers to a variety of symptoms that can be used to diagnose disease [13]. Many researchers have studied and analysed several techniques for predicting heart disease, including the following:

In [14], the most crucial features from four coronary artery disease (CAD) datasets were chosen using a novel heterogeneous hybrid feature selection technique. They also unveiled the Nasarian CAD dataset, a brand-new CAD dataset designed to evaluate the relationship between CAD and work-related characteristics. The efficiency of the suggested heterogeneous hybrid feature selection strategy is demonstrated by their findings.

In [15], a smart healthcare framework that improves the survival prognosis of heart failure patients without considering human feature engineering was proposed. Cloud computing and Internet of Things (IoT) technologies are used in this framework. The recommended method is to investigate whether heart failure patients can be classified as alive or dead using deep learning algorithms. The framework makes use of Internet of Things (IoT) sensors to gather data and transfer it to a cloud-based web server for analysis. There were 13 characteristics. The CNN model fared better than rival deep learning and machine learning models according to the testing data.

In [16], an approach for the detection of heart disease utilizing a feature choice optimization algorithm was reported, primarily focused on improving feature selection and minimizing the quantity of characteristics, and a recursive expansionist competitive method was used to choose relevant aspects of heart disease.

In [17], several machine learning and deep learning methods were employed to compare the findings of the UCI machine learning heart disease dataset, which comprises 14 key features utilized for the analysis. combined with a few multimedia tools, including portable electronics. An accuracy of 94.2% was attained using the deep learning method.

In [18], a new ensemble model called "NE-nu-SVC (Nested Ensemble nu-SVC) was introduced for the detection of CAD. Z-Alizadeh and Sani, two well-known CAD datasets, were used to test the proposed model. In [19], a multifilter approach was employed to increase the performance of different decision trees (DTs), which were subsequently applied to the CAD dataset. The power of DTs for CAD classification has been clearly discussed. In [20], a new training method for CAD datasets called the N2 genetic optimizer, which is based on genetics, was developed. For the classification step, three different SVMs (nuSVM, SVC, and LinSVM) were employed. The obtained outcomes indicated that the new technique outperformed other commonly used methods for CAD classification.

In [21], a study was conducted to increase prediction accuracy using different characteristic selection techniques, such as decision trees, naive Bayes, and neural network techniques, to predict cardiovascular disease or heart disease. It was found that the decision tree was accurate, scoring 98.54% in comparison with others. In light of this, a hybrid HRFLM method was proposed that combines the advantages of random forest (RF) and linear methods (LM), yielding an 88.4% prediction accuracy.

In [22], a study was conducted to increase the prediction accuracy by using different characteristic selection techniques and other techniques, such as decision trees, logistic regression, SVM, naive Bayes, and random forest. The results showed that logistic regression, with a score of 84.85%, was the best way to predict heart disease.

In [23], the authors used prediction models using different categories of characteristics, seven classification techniques (i.e., K-NN, DT, NB, LR, SVM, NN, and VOTE), and a hybrid method (i.e., logistic regression and naive Bayes). The outcomes demonstrated that the VOTE, along with the NB and LR methods, was the most accurate way to predict heart disease, with a score of 87.4%.

In [24], the multilayer Pi sigma neuron model (MLPSNM), which is based on the PI-sigma model, was presented for the aim of diagnosing cardiology using a machine learning repository. They utilized the cardiology dataset to do this. To minimize the dataset's dimensions and facilitate network learning, the BP algorithm was employed in conjunction with PCA and LDA preprocessing

methods. The grid converges after 50 iterations when using the SVM-LDA approach, which selects the features that are closest to the farthest level. In classifying patients with heart disease, the suggested model had a 94.53% classification accuracy.

In [25], an innovative transformer concept based on self-attention was developed to enhance the prediction of heart disease. The model attained a high accuracy of 96.51% when evaluated on the Cleveland dataset, surpassing the performance of other baseline approaches. By combining self-awareness systems and transformer networks, the model effectively captured contextual information and complex patterns in the data. The self-attention layers provided interpretability by assigning attention weights to different components of the input sequence, enabling physicians to understand the features that enhanced the forecasts of the model. The results show the potential of the proposed model for the early detection and diagnosis of cardiovascular diseases.

In [26], researchers proposed the CIGT format as a new way to integrate clinical, genetic, and patient transcriptome data with CVD data into a dataset suitable for artificial intelligence and machine learning (AI/ML). They then used for statistical tests to identify significant differences between patients and healthy individuals in terms of gene expression levels and clinical characteristics. Next, five different AI/ML classifiers were applied to predict the CVD status of patients based on their vital profiles. This study explored the discovery and prediction of biomarkers associated with CVD with high accuracy using a new combination of artificial intelligence and machine learning methods for precision medicine. This study presented a new approach that combines traditional statistics with a combination of artificial intelligence and machine learning techniques to identify important biomarkers from gene expression data of CVD patients and healthy individuals. This method revealed 18 tissue-specific biomarkers that can be used with up to 96% accuracy in disease prediction. Some of these biomarkers were previously known to be associated with CVD, while others were discovered for the first time. These biomarkers offer a useful foundation for identifying people at risk based on their biological profiles and may be useful indicators for the early diagnosis of CVD. Finally, they analysed transcriptomic data to validate the biomarkers discovered and understand their role in the course of the disease.

In [27], researchers aimed to develop accurate and efficient predictive models for the early detection of cardiovascular diseases using machine learning and deep learning techniques. Therefore, they used two different datasets to analyse the risk factors and features associated with cardiovascular diseases, namely, the cardiac heart disease dataset and the Cleveland heart disease dataset. Then, they implemented seven classifiers from machine learning and deep learning, namely, K-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR), convolutional neural network (CNN), gradient boosting (GB), XGBoost, and random forest (RF) classifiers. They evaluated and compared their performance using measures such as accuracy, sensitivity, specificity, and F1 score. They concluded that the XGBoost model is the best at achieving the highest levels of accuracy and reliability in predicting cardiovascular diseases, with 98.50% accuracy, 99.14% sensitivity, 98.29% specificity, and 98.71% F1 score.

In [28], the study aimed to create a machine learning model capable of predicting early-stage heart disease using different feature selection techniques. The UCI Cleveland dataset containing 303 heart patient records was used, and three feature selection methods were applied: ANOVA F value, chi-square test, and exchange of information. The random forest model outperformed six other machine learning models in predicting early-stage heart disease, with an accuracy of 94.51%, sensitivity of 94.87%, specificity of 94.23%, and AUROC of 94.95.

In [29], a general, hybrid framework for diagnosing heart problems using machine learning and data modelling techniques was presented. The framework used multiple feature selection and classification techniques, and the best result was determined using a new voting technique that considers classification probabilities. The framework is based on five feature selection techniques: Pearson correlation, analysis of variance, iterative elimination, soft regularization, and decision tree. It also uses six classification techniques: artificial neural networks, convolutional neural networks, random forests, logistic regression, and robust gradient boosting. Logistic regression is used as a second layer to vote on the results of the first layer. The framework uses a public dataset from the CAGL platform called the UC Irvine Heart Problems Dataset. This collection contains information on clinical cases of heart problems with 75 features and 76 columns plus nomenclature. The collection included four datasets provided by four healthcare facilities. The group used 13 basic features common to previous research. The study achieved an accuracy of up to 96.3% in diagnosing heart problems using the dataset used.

There are many studies related to the classification and prediction of heart disease. Several recent studies have used multiple algorithms and different data and machine learning algorithms, such as random forests, some of which are shown in Tables 1.

N-0	Study algorithms	Result	Accuracy
[25]	A new model based on self-attention and transformers	Predict heart disease using self-attention mechanisms and transformers	96.51%
[27]	Seven deep learning and machine learning classifiers	Identify heart problems	98.50%
[30]	Comparing the performance of three AutoML programs: AutoKeras, AutoGluon, and PyCaret	Predict heart disease using machine learning methods	86.89%
[31]	An automatic model capable of predicting heart disease with high accuracy	Predict heart disease	87.28%
[51]	Random forest ensemble classifier	Analyse how accurate the early coronary heart disease prognosis was.	89%
[52]	A random forest-based machine learning model	Enhancement and assess the effectiveness of the classifier	96%
[53]	Random forests	Increasing the precision with which coronary heart disease is classified	%85

Tables 1: Some Related Studies Predict Heart Disease

3. Materials and Methods

This section describes a new method based on a comparative study of techniques for classification and prediction. Moreover, this section discusses several important algorithms for machine learning and prediction that have been considered in the work and algorithm simulation of the Weka programme. A descriptive procedural analytical approach was used in this study to obtain data using different methods to collect and analyse the data and design a diagnostic system [32,33]. For identifying the most accurate techniques for diagnosing and classifying heart disease, 14 techniques, namely, RBF Network, LBR, Neural Network, Logistic Regression, Linear SVC, SVM, Random Tree, Naive Network by MLP Classifier algorithm, Random Forest, ID3, CART, J48, Decision Table, and Naive Bayes, were used and implemented through the WEKA program to simulate the correct decision-making process to achieve good accuracy in diagnosing and classifying heart disease. The study methodology involves collecting, studying, and analysing data; training on different classification techniques to select appropriate techniques for classifying data; and designing a web application to simulate the classification algorithm for diagnosing and classifying heart diseases, as shown in Figure 1.

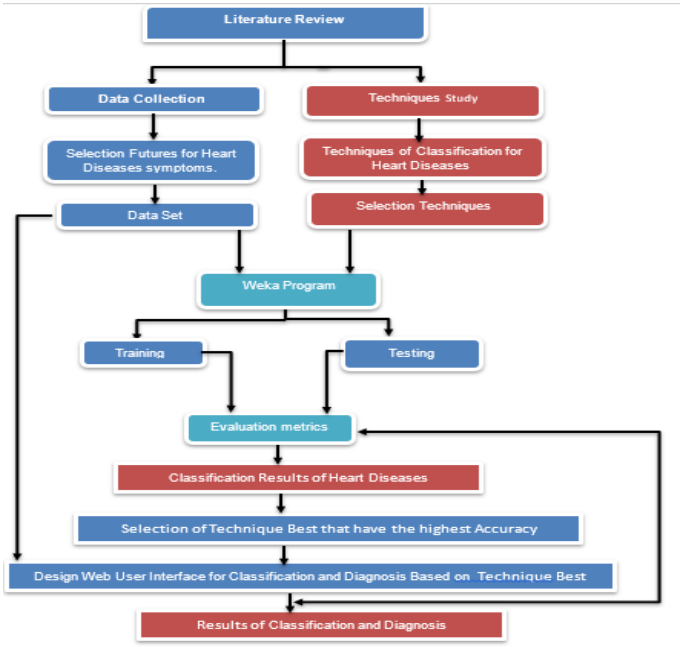

Figure 1: Methodology for Classification of Heart Diseases

a. Support Vector Machine (SVM)

SVM is a binary terminator with distinct properties. The work on SVMs can be summarized as follows. For example, we have an SVM training set that builds a super plane and a boundary between models in that the margin between positive points (class 1) and negative points (class 2) is as large as possible. It can divide and classify data into two categories—accepted or rejected—and can distinguish the boundaries between the data [34].

Unlike other models, the SVM model uses a kernel to separate data and calculates the cost and error rate by utilizing the following equation to estimate the similarity and difference between two inputs:

$$f(x) = \min_{0} C\frac{1}{m}\left[\sum_{i=1}^{m} \cos 1(\theta^T x^i) + (1-y^i)\cos 0(\theta^T x^i)\right] + \frac{1}{2}\sum_{j=1}^{n} \theta^2 \quad (1)$$

b. Id3 algorithm

The id3 algorithm has the divide and conquer principle; it is based on the idea of dividing the problem into parts. Every party solved the problem several times, and then the solutions were gathered. It is the optimal way for choosing the best feature or property [37] [38].

c. Logistic Regression

separates the data of a set of items into many sections based on comparable features. The error rate is computed based on the input and is determined using the logistic regression technique [40] [41] [54].

The error rate calculation function is:

$$f(x) = -y\log\frac{1}{1+e^{-\theta tx}} - (1-y)\log(1-\frac{1}{1+e^{-\theta tx}}) \quad (2)$$

d. Neural Network

It is a technique that analyses data that closely fit data with many attributes to obtain certain outcomes. There are many algorithms that work in this relation, such as the following:

• Kohenin network algorithm:

The algorithm follows these steps:

Activate the network, assume that the weight values are Wij (t) in the range 0 <= i <= n-1). These weights should be from element i to element j at time t as follows:

- Few random numbers for several inputs (n) are input for each arithmetic element.

The vicinity area around element j is set to include a large area, and this primary vicinity is denoted by Nj (0).

-The input values are set as shown in this equation:

$$(t), X_1(t), X_2(t), \ldots, X_{n-1}(t) = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots, -\infty < x < \infty \quad (3)$$

-where j(t) is the entry value of the input element (i) at time t.

-The distance dj between the input and each output element j is calculated as follows:

$$d_{i=\sum_{i=0}^{n-1}(x_i\ (t)-w_{ij}(t))^2} \quad (4)$$

The minimum distance is set, and the output element located at this distance is set to be j *

-The element j* weights as well as all the elements in the vicinity containing this winning element, which is symbolized by the symbol Nj* (t), are set to obtain new weights, as follows:

$$W_{ij}\ (t+1) = W_{ij}\ (t) + \eta(t)(X_i\ (t) - W_{ij}\ (t)) \quad (5)$$

-For element j in the vicinity of Nj* (t), with (0= i = <= n-1), where ŋ (t) is a gain factor. Its value is between (zero) and (one), and its value decreases with every adjustment circle of weights. Notably, the vicinity circle area decreases to include the least number of elements that are similar to and match the data and features of a particular input so that it is possible to create similar and active vicinity [42, 43].

e. Forest Random

It is a powerful and flexible algorithm in the field of machine learning that provides good results even without adjusting its parameters. It is one of the most widely used algorithms due to its ease and ability to be applied to regression and classification problems. This algorithm, as its label "Random Forest" suggests, creates a random forest. One of the most important advantages of RF is that it can be used for the classification and regression problems that make up most of today's machine learning problems [44]. Fig2 shows how RF could be applied to three trees whose results can be useful for the final output.

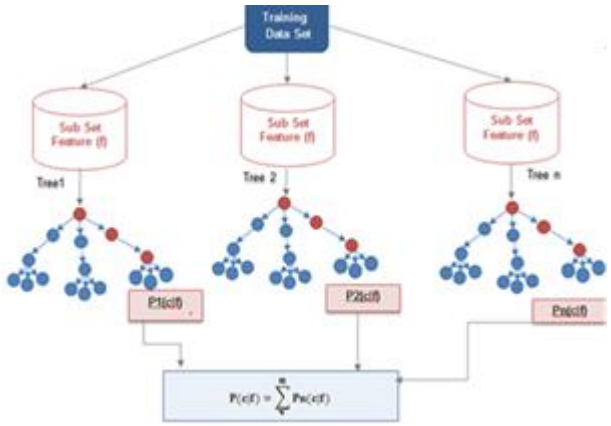

Figure 2: Random Forest Model with Three Trees

The random forest parameters are similar to those of the decision tree and bagging methods. However, the random forest method takes only a subspace of attributes into account when splitting a node. Trees can be more random when looking for the importance of the attribute randomly [45].

f. Decision Tree Rules

The path of a decision tree, which begins at any symptomatic node, continues via another symptomatic node, and finishes at the node

indicating the pathological condition, was used to categorize heart disease symptoms and pinpoint the primary symptoms underlying the disease [42, 50]. It continues with diagnosing further ailments. A selection of the decision tree rules utilized to classify various disease stages are shown in Table 2 :

R_N0	IF the patient:	The patient has:
1	Suffer from (Pain in chest (how pain (as pang at simple effort) and time(2 m-15 m)and location(on the top part of chest)) and or factors of dangerous(smoke or imbalance of blood pressure) and(dyspnea(at sleep or at simple effort)). or (tired and exhausted). and (heart throbbing(tachycardia)). and(cough(with phlegm)). and (swelling(parties)). and (pain in the right side and yellowing in eyes).	Congestive Heart Failure
2	Suffer from: - (Pain in chest (as pang location ((made of chest)). or (Factors of the dangerous(bacterial infection)). and(Dyspnea (at lying down or at doing task)). and(tired and exhausted). and(Heart throbbing (as acceleration)). or(Vertigo or spell) . and(Fever). and(Rash). and(Cutter pain).	Rheumatism in heart .
3	suffer from (Pain in chest (how pain (as pang move pain with effort)). or (factors of dangerous (smoking or old years or family history or bacterial infection)). and(dyspnea(simple effort)) . and (tired and exhausted). and(heart throbbing(tachycardia)). and(cough(with blood)). or(swelling (parties or legs)).	Valvular disease.
4	Suffer from: - (Pain in chest (as stress the chest , increasing by move and decrease by rest) and time (from 2 to 15 minutes)and location(left side or back of chest's bone)). and (Factors of the dangerous (smoking or fatness or old years or imbalance of blood pressure)). or (Dyspnea (at simple efforts)). or (tired and exhausted). or (Heart throbbing (as acceleration)). and (Vital wreaking (memory weak or seeing weak or move week)).	Arteriosclerosis

Table 2: Selection of the Decision Tree Rules

g. User Interface for Diagnosis

To categorize and forecast heart disease, the research data were applied using the classification algorithms used by Weka software. Following the procedure of classification, the dataset was applied to a web application created in the ASPX language, allowing us to replicate the proper method in this research by inputting data on the symptoms of the illness selected by the patient and forecasting his medical status. The method of the proposed system is shown inFig.3, which shows the working diagram of the diagnostic system.

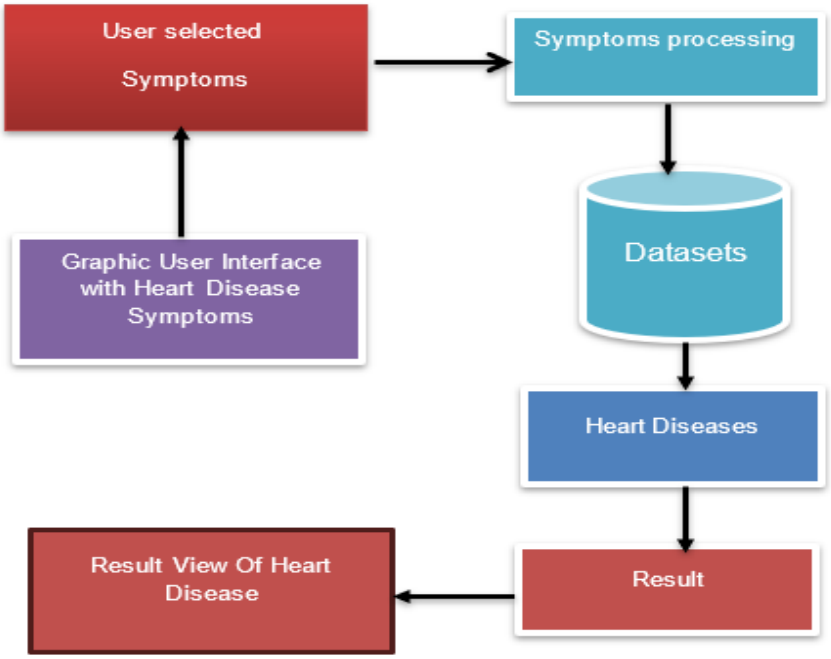

Figure 3: Diagram of the Diagnostic System

4. Results and Discussion

These techniques were tested using the Weka program to select a technique to simulate heart disease via the web application to predict and diagnose heart disease and obtain the best method for diagnosis. The data were subjected to different techniques in three ways: the first method involved entropy calculations and decision tree construction; the second technique required the use of the Weka program; and the third method used a web application to simulate the technology selected for the purpose of diagnosing heart disease. Most factors, techniques and methods were used to provide the correct result.

4.1 Entropy Calculation

Data distortion in the training set occurs when using the entropy equation, which is a method used to determine the root node and subsequent nodes using decision trees [46]. The entropy is calculated through the following equation:

$$Entropy = \sum_{i=1}^{c} -p_i \log_2 p_i \qquad (6)$$

Next, the gain is computed; finding the greatest benefit and dividing it by the overall entropy is the aim. The following formula is used to compute it:

$$Gain(p_i, i) = Entropy(p_i) - Entropy(i/p_i) \qquad (7)$$

4.2 Performance Metrics

The following generic equations reflect the general computations required to display the classifier's accuracy scores for the chosen method [47][48]:

$$TA = \frac{TP + TN}{TP + TN + FP + FN} \qquad (8)$$

$$RA = \frac{(tp+fp)*(tn+fn)*(fn+tp)*(FP+tp)}{Total^2} \qquad (9)$$

• The entropy of the instances when the characteristic (chest pain) appears, disease classification and number of pathological cases are calculated. We start with the first attribute in the Table.3, namely, chest pain. Four patients suffered from arteriosclerosis.

Disease Classification	N of Pathological Cases	Entropy of Chest pain
Arteriosclerosis	4	0
Arrhythmia	4	0
Heart Rheumatism	5	0
Angina Pectoris	4	0
Valvular disease	5	0
Heart contamination	6	0
Heart Rheumatism	4	0.556
Cardiomyopathy	4	
Arrhythmia	4	
Congenital defects	1	
Congenital defects	3	0.411
Heart Rheumatism	3	
Arrhythmia	1	

Table 3: Calculation of the Entropy of Chest Pain

Then, the gain (chest pain) attribute was calculated with Equation 7:

Gain(pi,i) = (4/46 * 0 + 4/46 * 0 + 5/46 * 0 + 4/46 * 0 + 5/46 * 0 + 6/46 * 0 + 13/46 * 0.556 + 5/46 * 0.411) = 0.785

The same method of calculating entropy is used for every characteristic. After that, each attribute is sorted, and the attribute with the highest entropy value located in the tree root is chosen. In the decision tree, ID3 was selected for the Weka system, and the decision tree was constructed as follows:

• ID3 Decision Tree

Acroanaesthesia = None
| Vital_weakness_on_the_bodys_parts = Yes
| | Palpitation = Yes
| | | Amnesia = no: arteriosclerosis
| | | Amnesia = Yes: Cardiomyopathy
| | Palpitation = No: Cardiomyopathy
| Vital_weakness_on_the_bodys_parts = No
| | Danger Factors = Yes
| | | Sickliness = No
| | | | Eruption = No
| | | | | Tumefaction = No
| | | | | | Dizziness = No
| | | | | | | Flatulence = No
| | | | | | | | Cough = No: Angina_pectoris
| | | | | | | | Cough = Yes: Valvular_disease
| | | | | | | Flatulence = Yes: Heart_contaminations
| | | | | | Dizziness = Yes
| | | | | | | Asthenia = No: Cardiomyopathy
| | | | | | | Asthenia = Yes: Myocardial_infraction
| | | | | Tumefaction = Yes
| | | | | | Headache = No
| | | | | | | Dyspnea = No: Congestive_heart_failure
| | | | | | | Dyspnea = Yes
| | | | | | | | Flatulence = No
| | | | | | | | | Cough = No: Congestive_heart_failure
| | | | | | | | | Cough = Yes
| | | | | | | | | | Asthenia = No: Valvular_disease
| | | | | | | | | | Asthenia = Yes: Congestive_heart_failure
| | | | | | | | Flatulence = Yes: Congestive_heart_failure
| | | | | | Headache = Yes
| | | | | | | Dyspnea = No: Valvular_disease
| | | | | | | Dyspnea = Yes: Heart Contaminations
| | | | Eruption = Yes
| | | | | Asthenia = No
| | | | | | Dyspnea = No: Rheumatism_in_heart
| | | | | | Dyspnea = Yes
| | | | | | | Palpitation = Yes: Heart Rheumatism
| | | | | | | Palpitation = No: Congestive_heart_failure
| | | | | Asthenia = Yes: Heart Contaminations
| | | | Rash = z: Valvular_disease
| | | Sickliness = Yes: Myocardial_infraction
| | Danger Factors = No
| | | Rhagades = No
| | | | Fever = No: Cardiomyopathy
| | | | Fever = Yes: Heart Rheumatism

| | | Rhagades = Yes: Congenital_defect
Acroanaesthesia = Yes: arrhythmia.

WEKA software was used to classify the research data using a variety of methods. Finally, GUIs were created to mimic random forest techniques so that users could efficiently identify heart illnesses using graphical user interfaces, which can also be used to diagnose other diseases.

4.3 Findings

The following findings were revealed from the results:

- The obtained data were used to design a diagnostic system for diagnosing common heart conditions.
- Various heart diseases can be classified as shown in Table (2).
- The classification using random forest was better than the other classifications, as shown in Table.4 below:

Algorithm Name	Accuracy
Bagging	80.7339%
LBR	76.1468%
RBFNetwork	87.156 %
RandomForest	**92.6606%**
RandomTree	88.9908%
SVM	72%
LinearSVC	86.3%
Logistic Regression	63.6%
Decision Table	68.8073%
J48	77.0642%
CART	77.981%
ID3	90.8257%
Naive Bayes	78.8991%

Table 4: The Results of Classification

- All the classification algorithms provided relatively correct answers.
- The random forest algorithm ranked the most highly at 92.66%, followed by ID3, as shown in Fig 4.
- Tables 1 shows that machine learning methods were used to classify and predict heart diseases for one type of heart disease, and the accuracy rate differed among the studies. For example, in [27][51][52][53], the random forest classifier was used, and the highest accuracy for classification was 98%. In the proposed study, the percentage was 92.66% for eight types of diseases.

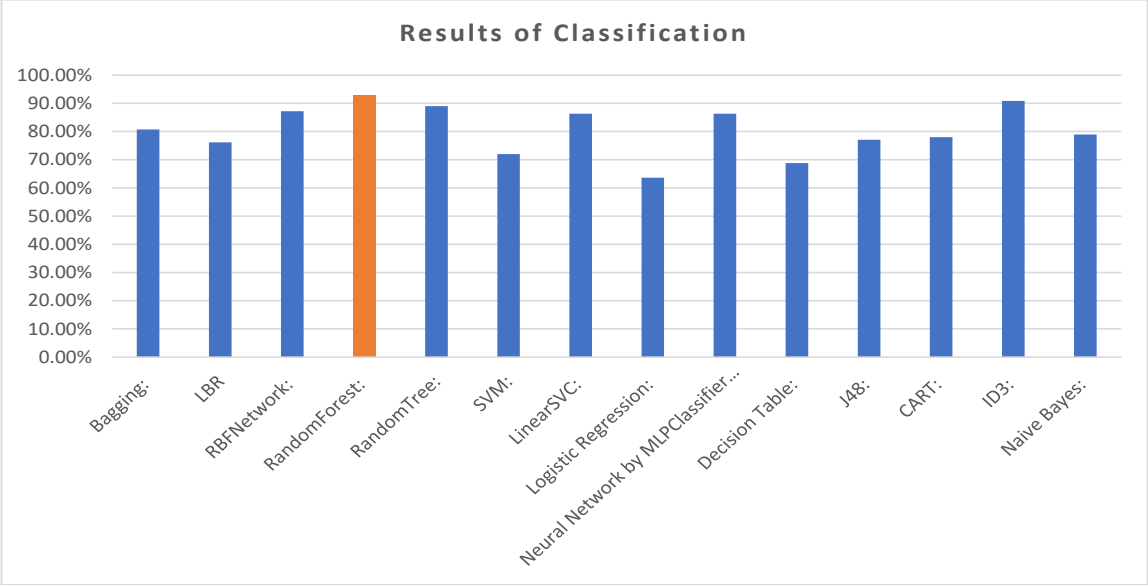

Figure 4: Values of Classification Accuracy for Heart Diseases

4.4 Conclusion and Future Work

Several machine learning approaches for categorizing various symptoms and heart conditions have been investigated, and an automated system for identifying heart conditions to support physicians in diagnostic clinics has been developed. The Weka tool was specifically used to apply 14 classifiers to imitate the proper decision-making process and achieve high accuracy in the diagnosis and classification of heart disease. The necessary tests were performed to determine how well the classification methods classified cardiac disorders. The results demonstrated that all the classification methods are predictive and capable of providing a reasonably accurate response. However, among all grading scales in the dataset, random forest was the most common, followed by the ID3 algorithm. The algorithm selected throughout the classification phase was simulated using a graphical user interface created in the ASPX language.

To improve healthcare and save expenses, more studies are needed to create a categorization system for methods and algorithms that will help identify the most suitable and efficient technology for a variety of ailments. Therefore, the usefulness of the existing search might be significantly increased by such a study. For instance, research may be performed on conditions that share biological traits with heart disease. It is also feasible to link the categorization system to additional systems that consider the primary signs and symptoms of the illness, such as an ECG system and temperature and heart rate monitoring. These systems may be combined to provide a full working environment.

References

1. Nalluri, S., Vijaya Saraswathi, R., Ramasubbareddy, S., Govinda, K., & Swetha, E. (2020). Chronic heart disease prediction using data mining techniques. *Advances in Intelligent Systems and Computing, 1079.*
2. Javeed, A., Khan, S. U., Ali, L., Ali, S., Imrana, Y., & Rahman, A. (2022). Machine learning-based automated diagnostic systems developed for heart failure prediction using different types of data modalities: A systematic review and future directions. *Computational and Mathematical Methods in Medicine, 2022.*
3. Bashir, S., Almazroi, A. A., Ashfaq, S., Almazroi, A. A., & Khan, F. H. (2021). A knowledge-based clinical decision support system utilizing an intelligent ensemble voting scheme for improved cardiovascular disease prediction. *IEEE Access, 9.*
4. Khaing, H. W. (2011). Data mining based fragmentation and prediction of medical data. In *Proceedings of the 2011 3rd International Conference on Computer Research and Development (ICCRD), 2*.
5. Shorewala, V. (2021). Early detection of coronary heart disease using ensemble techniques. *Informatics in Medicine Unlocked, 26.*
6. Tarawneh, M., & Embarak, O. (2019). Hybrid approach for heart disease prediction using data mining techniques. In Lecture Notes on Data Engineering and Communications Technologies (Vol. 29).
7. Singh, P., Singh, S., & Pandi-Jain, G. S. (2018). Effective heart disease prediction system using data mining techniques. *International Journal of Nanomedicine, 13*.
8. Hasan, O. S., & Saleh, I. A. (2021). Development of heart attack prediction model based on ensemble learning. *Eastern-European Journal of Enterprise Technologies, 4*(2–112).
9. Chung, K., Cho, H. Y., Kim, Y. R., Jhung, K., Koo, H. S., & Park, J. Y. (2020). Medical help-seeking strategies for perinatal women with obstetric and mental health problems and changes in medical decision making based on online health information: Path analysis. *Journal of Medical Internet Research, 22*(3).
10. Mohsen, A. A., Alrashahy, T., Naoufel, K., & Noaman, S. (2019). Use of comparative classification techniques to build a system for diagnosing heart diseases. In *Proceedings of the 2019 First International Conference of Intelligent Computing and Engineering (ICOICE)* (pp. 1-10).
11. Premsmith, J., & Ketmaneechairat, H. (2021). A predictive model for heart disease detection using data mining techniques. *Journal of Advances in Information Technology, 12*(1).
12. Odilbekov, F., Armonienė, R., Henriksson, T., & Chawade, A. (2018). Proximal phenotyping and machine learning methods to identify septoria tritici blotch disease symptoms in wheat. *Frontiers in Plant Science, 9.*
13. Junaid, M. J. A., & Kumar, R. (2020). Data science and its application in heart disease prediction. In *Proceedings of the International Conference on Intelligent Engineering and Management (ICIEM 2020).*
14. Nasarian, E., Abdar, M., Fahami, M. A., Alizadehsani, R., Hussain, S., Basiri, M. E., Zomorodi-Moghadam, M., Zhou, X., Pławiak, P., Acharya, U. R., Tan, R. S., & Sarrafzadegan, N. (2020). Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach. *Pattern Recognition Letters, 133.*
15. Umer, M., Sadiq, S., Karamti, H., Karamti, W., Majeed, R., & Nappi, M. (2022). IoT based smart monitoring of patients with acute heart failure. *Sensors, 22*(7).
16. Nourmohammadi-Khiarak, J., Feizi-Derakhshi, M. R., Behrouzi, K., Mazaheri, S., Zamani-Harghalani, Y., & Tayebi, R. M. (2020). New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection. *Health and Technology, 10*(3).
17. Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S., & Singh, P. (2021). Prediction of heart disease using a combination of machine learning and deep learning. *Computational Intelligence and Neuroscience, 2021.*
18. Aouabed, Z., Abdar, M., Tahiri, N., Champagne Gareau, J., & Makarenkov, V. (2020). A novel effective ensemble model for early detection of coronary artery disease.
19. Abdar, M., Nasarian, E., Zhou, X., Bargshady, G., Wijayaningrum, V. N., & Hussain, S. (2019). Performance

improvement of decision trees for diagnosis of coronary artery disease using multi filtering approach. 2019 *IEEE 4th International Conference on Computer and Communication Systems, ICCCS 2019.*

20. Abdar, M., Książek, W., Acharya, U. R., Tan, R. S., Makarenkov, V., & Pławiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. *Computer methods and programs in biomedicine, 179*, 104992.
21. Tarawneh, M., & Embarak, O. (2019). Hybrid approach for heart disease prediction using data mining techniques. In *advances in internet, data and web technologies: the 7th international conference on emerging internet, data and web technologies (EIDWT-2019)* (pp. 447-454). Springer International Publishing.
22. Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019, January). Improving heart disease prediction using feature selection approaches. In *2019 16th international bhurban conference on applied sciences and technology (IBCAST)* (pp. 619-623). IEEE.
23. Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics, 36,* 82-93.
24. Burse, K., Kirar, V. P. S., Burse, A., & Burse, R. (2019). Various pre-processing methods for neural network based heart disease prediction. In *Smart Innovations in Communication and Computational Sciences: Proceedings of ICSICCS-2018* (pp. 55-65). Springer Singapore.
25. Rahman, A. U., Alsenani, Y., Zafar, A., Ullah, K., Rabie, K., & Shongwe, T. (2024). Enhancing heart disease prediction using a self-attention-based transformer model. *Scientific Reports, 14*(1), 514.
26. DeGroat, W., Abdelhalim, H., Patel, K., Mendhe, D., Zeeshan, S., & Ahmed, Z. (2024). Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine. *Scientific Reports, 14*(1), 1-11.
27. Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M., & Qasem, S. N. (2024). Machine learning-based predictive models for detection of cardiovascular diseases. *Diagnostics, 14*(2), 144.
28. Biswas, N., Ali, M. M., Rahaman, M. A., Islam, M., Mia, M. R., Azam, S., Ahmed, K., Bui, F. M., Al-Zahrani, F. A., & Moni, M. A. (2023). Machine learning-based model to predict heart disease in early stage employing different feature selection techniques. *BioMed Research International, 2023,* Article ID 6864343.
29. Menshawi, A., Hassan, M. M., Allheeib, N., & Fortino, G. (2023). A hybrid generic framework for heart problem diagnosis based on a machine learning paradigm. *Sensors, 23*(3), 1392.
30. Paladino, L. M., Hughes, A., Perera, A., Topsakal, O., & Akinci, T. C. (2023). Evaluating the performance of automated machine learning (AutoML) tools for heart disease diagnosis and prediction. AI, 4(4), 1036–1058.
31. Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms, 16*(2), 88.
32. Abdulaziz Mohsen, A., Alsurori, M., Aldobai, B., & Abdulaziz Mohsen, G. (2019). New approach to medical diagnosis using artificial neural network and decision tree algorithm: Application to dental diseases. *International Journal of Information Engineering and Electronic Business, 11*(4).
33. Mansour, R. F., Amraoui, A. el, Nouaouri, I., Diaz, V. G., Gupta, D., & Kumar, S. (2021). Artificial intelligence and Internet of Things enabled disease diagnosis model for smart healthcare systems. *IEEE Access, 9.*
34. Dietterich, T. G. (2000). Ensemble methods in machine learning. In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 1857 LNCS.
35. Karaolis, M. A., Moutiris, J. A., Hadjipanayi, D., & Pattichis, C. S. (2010). Assessment of the risk factors for coronary heart events based on data mining with decision trees. *IEEE Transactions on Information Technology in Biomedicine, 14*(3).
36. Byliński, H., Sobecki, A., & Gebicki, J. (2019). The use of artificial neural networks and decision trees to predict the degree of odor nuisance of postdigestion sludge in the sewage treatment plant process. *Sustainability (Switzerland), 11*(16).
37. Wang, Y., Li, Y., Song, Y., Rong, X., & Zhang, S. (2017). Improvement of ID3 algorithm based on simplified information entropy and coordination degree. *Algorithms, 10*(4).
38. Yang, S., Guo, J. Z., & Jin, J. W. (2018). An improved ID3 algorithm for medical data classification. *Computers and Electrical Engineering, 65.*
39. Parameswari, D., & Khanaa, V. (2020). Intrusion detection system using modified J48 decision tree algorithm. *Journal of Critical Reviews, 7*(4).
40. Makaryus, A. N., Makaryus, J. N., Figgatt, A., Mulholland, D., Kushner, H., Semmlow, J. L., Mieres, J., & Taylor, A. J. (2013). Utility of an advanced digital electronic stethoscope in the diagnosis of coronary artery disease compared with coronary computed tomographic angiography. *American Journal of Cardiology, 111*(6).
41. Ahamed, B. S. (2021). Prediction of type-2 diabetes using the LGBM classifier methods and techniques. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12*(12), 223-231.
42. Kim, J. K., & Kang, S. (2017). Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis. *Journal of healthcare engineering, 2017*(1), 2780501.
43. Desai, F., Chowdhury, D., Kaur, R., Peeters, M., Arya, R. C., Wander, G. S., Gill, S. S., & Buyya, R. (2022). HealthCloud: A system for monitoring health status of heart patients using machine learning and cloud computing. *Internet of Things (Netherlands), 17.*
44. El-Shafiey, M. G., Hagag, A., El-Dahshan, E. S. A., & Ismail,

M. A. (2022). A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest. *Multimedia Tools and Applications, 81*(13), 18155-18179.

45. Huang, W., Ying, T. W., Chin, W. L. C., Baskaran, L., Marcus, O. E. H., Yeo, K. K., & Kiong, N. S. (2022). Application of ensemble machine learning algorithms on lifestyle factors and wearables for cardiovascular risk prediction. *Scientific Reports, 12*(1), 1033.
46. Heath, A., Gonzales, M., & von Alvensleben, I. (2019). Variable selection for early diagnosis of congenital heart disease using random forest entropy calculations. Cardiology in the Young, 29.
47. Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. *IEEE Access, 8,* 107562–107582.
48. Abdar, M., Zomorodi-Moghadam, M., Das, R., & Ting, I. H. (2017). Performance analysis of classification algorithms on early detection of liver disease. *Expert Systems with Applications, 67,* 239-251.
49. Alaiad, A., Najadat, H., Mohsen, B., & Balhaf, K. (2020). Classification and association rule mining technique for predicting chronic kidney disease. *Journal of Information & Knowledge Management, 19*(01), 2040015.
50. Quancheng, Z., & Jingbin, H. (2021). Research on data mining of physical examination for risk factors of chronic diseases based on classification decision tree. In *2021 IEEE 6th International Conference on Intelligent Computing and Signal Processing (ICSP).*
51. Ani, R., Augustine, A., Akhil, N. C., & Deepa, O. S. (2016). Random forest ensemble classifier to predict coronary heart disease using risk factors. In L. Suresh & B. Panigrahi (Eds.), Proceedings of the International Conference on Soft Computing Systems (pp. 773-782). Advances in Intelligent Systems and Computing, vol 397. Springer, New Delhi.
52. Jawalkar, A. P., Swetcha, P., Manasvi, N., et al. (2023). Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting. *Journal of Engineering and Applied Sciences, 70,* 122.
53. Wang, J., Rao, C., Goh, M., et al. (2023). Risk assessment of coronary heart disease based on cloud-random forest. *Artificial Intelligence Review, 56*, 203–232.
54. Al-Majmar, N. A., Mohsen, A. A., & Al-Thulathi, M. S. (2022). Development a Model for Drug Interaction Prediction Based on Patient State. *International Journal of Intelligent Systems and Applications, 13*(6), 28.