Classification and Diagnosis of Heart Disease Using Machine Learning

Ayedh Abdulaziz Mohsen
ayedh982001@hotmail.com

Ibb
Kharroubi Naoufel
Taif
Taher Alrashahy
Hodiadah
Somia Noaman
Ibb

Research Article

Keywords: Heart diseases, Classification, Diagnosis, Techniques, Machine learning

Posted Date: February 27th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3985932/v1

License: ☑️ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: The authors declare no competing interests.
Abstract: Heart disease is one of the most common causes of death worldwide, and accurate and rapid diagnosis is needed to avoid serious complications. This study aimed to present a system for diagnosing and classifying common heart diseases using machine learning techniques. The data used are a set of medical records for heart patients at Ibb Medical Clinic, Ibb City, Yemen. Fourteen different classification algorithms were implemented using the Weka tool to classify different heart diseases, including eight heart diseases. The results showed that the RF algorithm was the best in terms of accuracy, sensitivity, specification, and F-measure. Based on this algorithm, a web system based on the ASPX language was designed to help users enter patient data and obtain a diagnosis and classification of heart disease. The system can help doctors in the diagnosis process and increase the medical awareness of the community.

Keywords: Heart diseases, Classification, Diagnosis, Techniques, Machine learning.

I. Introduction

Currently, heart disease is one of the most prevalent diseases in the world. It has been estimated that it caused approximately 17.9 million deaths in 2017, i.e., 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., blood pressure, cholesterol, heart rate, and glucose level [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 in which the rate of occurrence is very high, especially in developing countries, 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 Naïve 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 are sensitive and accurate, but the latter depends on the degree of data accuracy stored 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 final result (e.g., programs that diagnose diseases based on symptoms). Because many doctors lack enough information about the features of diseases, computer systems, and interconnected technologies that help patients identify symptoms of diseases, they may not be able to diagnose diseases correctly in a short time, especially when patients suffer from more than one disease [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 records in hospitals and clinics 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 different types of heart diseases, such as artery occlusion, heart-related rheumatism, angina pectoris heart disease, heart-related disorders, birth defects, heart arrhythmia and cardiomyopathy. The structure of the article is organized as follows: In Section II, a discussion of relevant work in this area is included. Section III describes the materials, methods, and approach used. Section IV discusses the results and findings. Finally, Section V presents the conclusions of this study and provides some future work directions.

II. Related works

Before discussing studies that diagnose and classify heart disease, it is important to define heart disease. The heart is the most important organ in the human body because it pumps blood to all parts of the body[13]. The term "heart disease" refers to a variety of symptoms that can be used to diagnose heart disease. Many researchers have studied and analysed many techniques for predicting heart disease, including the following:

In [14], a new heterogeneous hybrid feature selection method was proposed for identifying the most important features among 4 coronary artery disease (CAD) datasets. In addition, they introduced a new CAD dataset called the Nasarian CAD dataset to assess the association between work-related features and CAD. Their results indicate the effectiveness of the proposed heterogeneous hybrid feature selection method.

In [15], a smart healthcare framework that improves the survival prognosis for 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 IoT-based sensors to collect signals and send them to a cloud-based web server for analysis. There were 13 characteristics included in this study. The experimental results showed that the CNN model outperformed competing deep learning and machine learning models.

In [16], an approach for the detection of heart disease utilizing a feature selection 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. The K-nearest neighbour method was also used for classification.

In [17], several machine learning methods and deep learning methods were employed to compare the findings and analysis 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. The proposed model was tested on two well-known CAD datasets: Z-Alizadeh and Sani. In [19], a multfilter approach was used to improve the performance of different decision trees (DTs) and then applied to the CAD dataset. The power of DTs for CAD classification has been clearly discussed. Abdar et al. [20] developed a new training method for CAD datasets called the N2 genetic optimizer, which is based on genetics. 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. Accordingly, a hybrid HRFLM approach was proposed by combining the characteristics of the linear method (LM) and random forest (RF), resulting in a prediction accuracy of 88.4%.

In [22], a study was conducted to increase the accuracy of prediction by applying 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 researchers 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 technique (i.e., logistic regression and naive Bayes). The results showed 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], a machine learning repository was used for diagnosing heart disease; for this purpose, they used the cardiology dataset and proposed the multilayer Pi sigma neuron model (MLPSNM) for diagnosing cardiology based on the PI-sigma model. Both architecture and computation are less complex than those of other models that were proposed earlier. To learn the network, the BP algorithm and PCA and LDA preprocessing methods were used to reduce the dimensions of the dataset. In the SVM-LDA method, the closest features to the furthest level are selected to check the grid, which converges after 50 iterations. The proposed model achieved a 94.53% classification accuracy for diagnosing heart disease.

In [25], a novel self-attention-based transformer model was developed to enhance the prediction of heart disease. The model achieved a high accuracy of 96.51% when evaluated on the Cleveland dataset, surpassing the performance of other baseline approaches. By combining self-attention mechanisms 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 contributed to the model's predictions. The findings demonstrate 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 transcriptomic data of patients with cardiovascular disease (CVD) into a dataset suitable for artificial intelligence and machine learning (AI/ML). They then used four 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 techniques 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 to predict disease with up to 96% accuracy. Some of these biomarkers were previously known to be associated with CVD, while others were discovered for the first time. These biomarkers may be potential indicators for the early detection of CVD and provide a valuable framework for identifying patients at risk based on their biological profiles. Finally, they analysed the 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). They evaluated and compared their performance using metrics 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 build 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 mutual information. Then, six different machine learning models were applied to find the most optimistic model with the best feature set. The results showed that the random forest model provided the best performance for the SF3 feature set, with an accuracy of 94.51%, sensitivity of 94.87%, specificity of 94.23%, AUROC of 94.95, and log loss of 0.31. The performance of the proposed model with the selected features indicates that it has high potential for clinical use in predicting early-stage heart disease at a cost.

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 takes into account 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: logistic regression, sigmoid support machine, random forest, robust gradient boosting, artificial neural network, and convolutional neural network. 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 76 columns and 75 features plus nomenclature. The collection included four different databases provided by four different medical institutions. 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.

<table>
<thead>
<tr>
<th>N-0</th>
<th>Study algorithms</th>
<th>Result</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>[25]</td>
<td>A new model based on self-attention and transformers</td>
<td>Predict heart disease using self-attention mechanisms and transformers</td>
<td>96.51%</td>
</tr>
<tr>
<td>[27]</td>
<td>Seven classifiers of machine learning and deep learning</td>
<td>Predict heart disease</td>
<td>98.50%</td>
</tr>
<tr>
<td>[30]</td>
<td>Comparing the performance of three AutoML tools - PyCaret, AutoGluon, and AutoKeras</td>
<td>Predict heart disease using machine learning techniques</td>
<td>86.89%</td>
</tr>
<tr>
<td>[31]</td>
<td>An automatic model capable of predicting heart disease with high accuracy</td>
<td>Predict heart disease</td>
<td>87.28%</td>
</tr>
<tr>
<td>[51]</td>
<td>Random forest ensemble classifier</td>
<td>Evaluate the accuracy of the early prediction of coronary heart</td>
<td>89%</td>
</tr>
<tr>
<td>[52]</td>
<td>A random forest-based machine learning model</td>
<td>Optimization and evaluate the classifier's performance</td>
<td>96%</td>
</tr>
</tbody>
</table>
III. 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 program. 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, RBFNetwork, LBR, Neural Network, Logistic Regression, LinearSVC, SVM, RandomForest, Naive Network by MLP Classifier algorithm, RandomForest, 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 Fig 1.

![Methodology for Classification of Heart Diseases](image)

**a. Support Vector Machine**

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].

---

| [53] | Random forests | Improving the classification accuracy of coronary heart disease | %85 |
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:

\[
f(x) = \min_{0} \frac{1}{m} \sum_{i=1}^{n} \left[ \cos(\theta x_i) + (1 - y_i) \cos(\theta y_i)^{2} \right] + \frac{1}{2} \sum_{i=1}^{n} \theta^{2}
\]  

(1)

b. Id3 algorithm

The id3 algorithm was used for all classification problems (income analysis, disease classification, image classification and other classification problems). It 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 solution was 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]. The error rate calculation function is:

\[
f(x) = -y \log \left( \frac{1}{1 + e^{tx}} \right) - (1 - y) \log \left( 1 - \frac{1}{1 + e^{tx}} \right)
\]  

(2)

d. Neural Network

It is a technique that analyses data that closely fit data with a large number of features to reach certain results. 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 \( W_{ij} \) in the range \( 0 \leq i \leq n-1 \). These weights should be from element i to element j at time t as follows:
  
  - Small random values for a number of 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 \( N_j \) (0).
  - The input values are set as shown in this equation:
    \[
    \theta^{(t)}, \theta^{(t)}, \ldots, \theta^{(t)} \rightarrow 1 + \frac{1}{2} \sum_{i=1}^{n} \theta^{(t)}
    \]  
    - Where \( j(t) \) is the entry value of the input element (i) at time t.
  - The distance \( d_j \) between the input and each output element j is calculated as follows:
    \[
    d_j = \sum_{i=2}^{n} (x_i - w_{ij})^2
    \]  
    - 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 \( N_{j^*} \) (t), are set to obtain new weights, as follows:
    \[
    W_{ij} \rightarrow W_{ij} + \eta(t)(x_i - w_{ij})
    \]  
    - For element j in the vicinity of \( N_{j^*} \) (t), with \( 0 \leq i \leq n-1 \), where \( \eta(t) \) 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 used for classification and regression 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]. Fig 2 shows how Forest Random could be applied to three trees whose results can be useful for the final output.

![Fig2. Random forest model with three trees](image)

The random forest parameters are similar to those of the decision tree and bagging algorithms. However, the Random Forest algorithm 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. Table 2 displays a few of the decision tree rules used to categorize different disease states:

<table>
<thead>
<tr>
<th>R_N0</th>
<th>IF the patient:</th>
<th>The patient has:</th>
</tr>
</thead>
</table>
| 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**      |
<table>
<thead>
<tr>
<th></th>
<th>Symptoms</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Suffer from: -&lt;br&gt;(Pain in chest (as pang location ((made of chest))).&lt;br&gt;or (Factors of the dangerous(bacterial infection)).&lt;br&gt;and(Dyspnea (at lying down or at doing task)).&lt;br&gt;and(tired and exhausted).&lt;br&gt;and(Heart throbbing (as acceleration)).&lt;br&gt;or(Vertigo or spell).&lt;br&gt;and(Fever).&lt;br&gt;and(Rash).&lt;br&gt;and(Cutter pain).</td>
<td>Rheumatism in heart.</td>
</tr>
<tr>
<td>3</td>
<td>Suffer from&lt;br&gt;(Pain in chest (how pain (as pang move pain with effort))).&lt;br&gt;or (factors of dangerous (smoking or old years or family history or bacterial infection)).&lt;br&gt;and(dyspnea(simple effort)).&lt;br&gt;and (tired and exhausted).&lt;br&gt;and(heart throbbing(tachycardia)).&lt;br&gt;and( cough(with blood)).&lt;br&gt;or(swelling (parties or legs)).</td>
<td>Valvular disease.</td>
</tr>
<tr>
<td>4</td>
<td>Suffer from: -&lt;br&gt;(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)).&lt;br&gt;and (Factors of the dangerous (smoking or fatness or old years or imbalance of blood pressure)).&lt;br&gt;or (Dyspnea (at simple efforts)).&lt;br&gt;or (tired and exhausted).&lt;br&gt;or (Heart throbbing (as acceleration)).&lt;br&gt;and (Vital wreaking (memory weak or seeing weak or move week)).</td>
<td>Arteriosclerosis</td>
</tr>
</tbody>
</table>

**g. User Interface for Diagnosis**

To categorize and forecast heart disease, the research data were applied using the classification algorithms used by the WEKA software. Following the classification process, 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 application method is shown in Fig.3, which shows an overview of the proposed diagnosis.
IV. 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 method involved the use of the Weka program; and the third method involved the use of 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.

- Entropy calculation
  Data distortion in the training set occurs when using the entropy equation, which is a method used to calculate the root node and subsequent nodes using decision trees [46]. The entropy is calculated through the following equation:
  \[
  \text{Entropy} = \sum_{i=1}^{c} -p_i \log_2 p_i
  \]  
  \(6\)
  Next, compute the gain; the goal is to obtain the largest gain and divide it by the total entropy. The following formula is used to compute it:
  \[
  \text{Gain}(p_i, i) = \text{Entropy}(p_i) - \text{Entropy}(i/p_i)
  \]  
  \(7\)

Performance Metrics
The following generic equations reflect the general computations required to display the classifier's accuracy scores for any chosen algorithm[47][48]:
  \[
  TA = \frac{TP + TN}{TP + TN + FP + FN}
  \]  
  \(8\)
Entropy calculation of the cases in which the attribute (chest pain) appears, disease classification and the number of pathological cases: We start with the first attribute in the Table 3, namely, chest pain. Four patients suffered from arteriosclerosis.

<table>
<thead>
<tr>
<th>Disease Classification</th>
<th>N of Pathological Cases</th>
<th>Entropy of Chest pain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arteriosclerosis</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Arrhythmia</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Heart Rheumatism</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Angina Pectoris</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Valvular disease</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Heart contamination</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Heart Rheumatism</td>
<td>4</td>
<td>0.556</td>
</tr>
<tr>
<td>Cardiomyopathy</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Arrhythmia</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Congenital defects</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Congenital defects</td>
<td>3</td>
<td>0.411</td>
</tr>
<tr>
<td>Heart Rheumatism</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Arrhythmia</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

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

$$\text{Gain}(p_i) = (4/46 \times 0 + 4/46 \times 0 + 5/46 \times 0 + 4/46 \times 0 + 5/46 \times 0 + 6/46 \times 0 + 13/46 \times 0.556 + 5/46 \times 0.411) = 0.785$$

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 that is located in the tree root is chosen. In the decision tree, ID3 was selected for the WEKA application, 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, GUls 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.

- Findings
The following findings were revealed from the results:
- The obtained data were used to design a web system for diagnosing common heart diseases.
- Different heart diseases can be classified as shown in Table (2).
- It was found that the classification using RF was better than the other classifications, as shown in Table.4 below:

<table>
<thead>
<tr>
<th>Algorithm Name</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bagging</td>
<td>80.7339 %</td>
</tr>
<tr>
<td>LBR</td>
<td>76.1468%</td>
</tr>
<tr>
<td>RBFNetwork</td>
<td>87.156 %</td>
</tr>
<tr>
<td>RandomForest</td>
<td>92.6606%</td>
</tr>
<tr>
<td>RandomTree</td>
<td>88.9908%</td>
</tr>
<tr>
<td>Algorithm</td>
<td>Accuracy</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>SVM</td>
<td>72%</td>
</tr>
<tr>
<td>LinearSVC</td>
<td>86.3%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>63.6%</td>
</tr>
<tr>
<td>Neural Network by MLPClassifier algorithm</td>
<td>86.3%</td>
</tr>
<tr>
<td>Decision Table</td>
<td>68.8073%</td>
</tr>
<tr>
<td>J48</td>
<td>77.0642%</td>
</tr>
<tr>
<td>CART</td>
<td>77.981%</td>
</tr>
<tr>
<td>ID3</td>
<td>90.8257%</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>78.8991%</td>
</tr>
</tbody>
</table>

- 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. Its application has proven to be scientifically efficient in diagnosing heart disease and determining the rate of infection depending on the number of symptoms inserted and other cases when showing the natural health of patients.

Table 1 shows that machine learning algorithms 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.

![Results of Classification](image-url)

**Fig 4.** Values of classification accuracy for heart diseases
V. Conclusion and Future Work

Several data mining approaches were investigated to categorize various symptoms and heart conditions, and an automated system to identify heart conditions was investigated to support physicians in diagnostic clinics. 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 algorithms 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 find 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 take into account 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:


