

Independent Study Final Report

Level 3

Heart Disease Prediction Using Machine Learning Algorithms

by

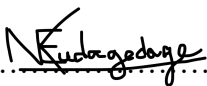
Kudagodage N.E.

Student

Signature

Date

Kudagodage N.E.


.....


13/11/2022
.....

Supervisor

Signature

Date

Dr.(Mrs.) Fernando K.S.D.


.....

13/11/2022
.....

Faculty of Information Technology

University of Moratuwa

2022

Heart Disease Prediction Using Machine Learning Algorithms

Kudagodage N.E.
Faculty of Information Technology
University of Moratuwa
nelmikudagodage@gmail.com

Abstract—Heart disease is becoming more common every day at an unprecedented and exponential rate. It has been the leading cause of death worldwide for the past few decades. To identify heart diseases early and treat them successfully, it is crucial to find a reliable and accurate method for automating the task. Every day, the health care sector produces enormous amounts of data about patients and diseases. Processing massive quantities of data in the medical field requires the use of data science. Machine learning algorithms and techniques have been used on a variety of medical datasets to automate the analysis of complex data sets. This study evaluates the comparative survey of ML classification methods proposed to assist medical professionals in heart disease diagnosis prediction. Begin by providing an overview of machine learning and then, it gives an overview of the work that has already been done and gives an understanding of the current used techniques.

Keywords—*Machine Learning, Heart Disease Prediction, cardiovascular disease prediction, Decision Tree, Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbors, Artificial Neural Network, UCI dataset*

I. INTRODUCTION

Cases of heart disease are increasing at an alarming rate, and it is crucial and concerning to anticipate any such diseases in beforehand. The World Health Organization estimates that more than 10 million people worldwide pass away each year as a result of heart disease. [1] Even any disturbance in the heart affects the entire body. The field of medical dictation has always required a lot of upkeep in terms of time, accurateness, and cost. People are fallible and prone to making mistakes. The human mind cannot handle too much estimation when pointing out any prediction based on numerous factors, so it may repeatedly provide incorrect feedback, causing a significant risk to the patient. Therefore, it is crucial to make the right prediction about the disease at the right time.

Data mining has the potential to uncover information from large datasets that have hidden patterns in the medical field. A personified structure is required for this data arrangement. Processing data is indeed a prerequisite for using machine learning techniques and obtaining more accurate outcomes. This extractive data will aid in the prediction of the medical diagnosis using machine learning techniques. With the aid of the dataset's historical patterns, a future prediction-based approach will also assist the physician in taking the appropriate actions to treat the patient quickly and efficiently. The accurate prognosis of the disease is made possible by ML techniques and prediction models. [2]

Heart disease encompasses a wide range of conditions that can harm your heart. Heart disease covers a wide range of illnesses, including congenital heart defects, blood vessel diseases like coronary artery disease, issues with heart rhythm (arrhythmias), and more. Heart disease and cardiovascular disease are sometimes used interchangeably. A myocardial infarction (heart attack), angina (chest pain), or a stroke can all be caused by blocked or narrowed blood vessels, which is referred to as having cardiovascular disease (CVD). 17.9 million people worldwide experience cardiovascular disease each year (CVD). With a mortality rate of more than 17.7 million per year, it causes nearly 32% of all deaths worldwide. According to WHO statistics, cardiovascular disease (CVD) accounts for 37% of all premature deaths globally, with low- and middle-income countries experiencing an overwhelming 82% of these deaths as a result of the slow and unreliable detection of heart disease.[3], [4] The prevalence of cardiovascular disease is steadily rising as a result of people's unhealthy lifestyle choices, such as smoking, eating a lot of fat, and being less physically active.

This review paper provides a comprehensive overview of the classification algorithms of ML applied throughout the research area of heart disease prediction and how they were applied by earlier researchers. It intends to shed light on the significance of machine learning around the health industry and demonstrates how it may support medical professionals by producing precise forecasts. Making decisions with discrete data is a challenging and complex task. Data mining's subfield of machine learning (ML) effectively manages large, well-organized datasets. The classification and diagnosis of cardiovascular diseases both might benefit from the application of machine learning techniques. Machine learning has a variety of applications, from the identification of factors that increase the risk of illness to the improvement of automotive safety systems. Machine learning offers the most prominent predictive modeling techniques to get around current limitations. This study examines the effectiveness of several machine learning (ML) methods, including Naive Bayes, Decision Trees, Logistic Regression, Random Forest, Support Vector Machines, Artificial Neural Network and K-Nearest Neighbors.

The remaining sections of the article are arranged as follows. Section two provides background information on machine learning, classification methods, and the most popular dataset for heart disease research. The existing proposed research study in this area is reviewed in section 3 along with a table form comparing of the classification algorithms discussed

in section 2 based on their accuracy. Last but not least, sections 4 and 5 present the discussion and conclusion.

II. BACKGROUND

Brief analyses of the topics are provided in this section, that are related to the main theme of the paper, including machine learning, its methods, data preprocessing, metrics for measuring performance, and a brief description of the most widely used cardiovascular disease dataset.

A. Machine Learning

Building algorithms that can be taught from experience is the focus of the artificial intelligence field known as machine learning (ML). ML algorithms work by uncovering disguised sequences in the input dataset and building model types based on those patterns. Then, they are able to make precise predictions for fresh datasets that the algorithms have never seen before. As a result, the machine learned and developed an increased level of intelligence, enabling it to recognize patterns that would be very challenging or impossible for a human to notice on their own. Large datasets can be processed by ML algorithms and techniques, which can also produce predictions. Therefore, it is a multidisciplinary research field that draws inspiration from numerous scientific and technical areas of study.

B. Machine Learning Algorithms

Unsupervised learning, supervised learning, semi-supervised learning, and reinforcement learning are the four main categories under which machine learning algorithms can be divided. The focus of the following section is on how each type of approach might be applied to solve real-world problems.[5]

1) Supervised Learning

This method makes use of a dataset that includes examples and the responses to them. During a training phase, the methodology can pick up knowledge from a dataset and apply it to new datasets as input. Regression and classification are two instances of supervised learning methods.[6]

2) Unsupervised learning

The responses for this technique are unavailable from the dataset. In order to categorize the input values, the algorithm attempts to identify similarities between them. This method, relying on data instead of human input, allows for the examination of unlabeled datasets. In most instances, this is done to draw out the generative properties, find significant patterns & structures, grouping in results, and explore purposes. Examples of typical unsupervised learning tasks include clustering and association rule creation.[6]

3) Semi-supervised Learning

Semi supervised learning may be considered a hybrid of the two kinds of learning methods previously discussed due to the use of both labeled and unlabeled data. When labeled data is in short supply and unlabeled data is plentiful, semi-supervised learning performs well. By utilizing only labeled data from the model, semi-supervised learning seeks to increase prediction accuracy over unsupervised learning. Semi-supervised learning has many applications, including translation, fraud detection, data labeling, and text classification.[5]

4) Reinforcement Learning

As it associates with the environment, this model gets better at what it does. learn how to fix its errors as a result. Through investigation and testing various options, it should obtain the desired outcome. An environment-driven strategy to machine learning allows software agents and computers to use reinforcement learning to instantly determine the best behavior in a specific situation or environment. [6]

Supervised learning is the most widely used learning approach, and the classification method in particular is frequently used for forecasting. This paper primarily evaluates studies that employed classification algorithms to early prediction of cardiovascular diseases.

C. Classification Machine Learning Techniques

Data models are categorized into desired classifications through the process of classification. Classification makes predictions for potential cases based on historical data. Based on the data points themselves, the classification approach predicts the target class for each data point. This section provides a brief definition of the most common classification methods for forecasting cardiovascular heart disease.

1) K-Nearest Neighbor (KNN)

A straightforward, user-friendly supervised machine learning algorithm known as the k-nearest neighbors (KNN) algorithm can be used to resolve classification and regression issues. The KNN algorithm presumes that similar things can be found nearby. Or to put it differently, things that are related seem to be nearby. KNN uses the data to categorize new data points according to the similarity metric.[4] KNN extracts the data points from the dataset and calculates the closest result. It offers extremely high predictive accuracy. The cardiovascular disease dataset contains a number of features, making this technique ideal for pattern recognition. The majority of KNN extracts knowledge and logic depending on the Euclidean distance Samples function $d(x_i, x_j)$. [3]

$$d(x_i, x_j) = \sqrt{(x_{i,1} - x_{j,1})^2 + \dots + (x_{i,m} - x_{j,m})^2} \quad (1)$$

2) Logistic regression

Logistic regression is used to predict the likelihood of a target variable when the dependent variable (target) is categorical. The outcome of the goal variable in logistic regression can differ in two ways. It calculates the likelihood that data's goal parameter will occur and suggests that the input and the output have a linear relationship. There are only two possible classes, 0 for failure and 1 for success, because the target or dependent variable is dichotomous in nature. The dependent variable is still seen by LR as a bi-categorical attribute. Its primary applications are in forecasting and success probability calculations. The only distinction from linear regression is that the variable's outcome is categorical rather than continuous. As a result, Logistic Regression uses the Sigmoid function, a cost function that is more complicated. The cost function's range in the model of logistic regression is 0 to 1. So the expectation of the Logistic Regression Hypothesis is:[7] [8]

$$0 \leq h_{\theta}(x) \leq 1 \quad (2)$$

The $h_{\theta}(x)$ represents the hypothesis expectation for any instance or data point x . and then used the sigmoid function to map the calculated probabilities. The sigmoid function can change any real value into a number between 0 and 1. The study predictive model is overseen and searches for a yes-or-no response. That is why this algorithm is used to forecast cardiovascular disease.[7]

3) Naive Bayes

We also use the Naive Bayes classifier, a machine learning technique, to categorize data and predict an instance's probability. Therefore, each Naive Bayes classifier makes the assumption that the value of a specific feature is unique from all the other features provided in the class variables. The Bayes rule is the foundation of the Nave Bayes algorithm. The primary presumption and most crucial factor in classifying a dataset is its attribute independence. It can be predicted quickly and easily, and it performs better whenever the independence presumption seems to be true. The following is the classifier that resulted from the Bayes theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (3)$$

By applying the Bayes theorem from the above equation, we can determine the likelihood that instance A will occur given that instance B has already occurred. Here, we can consider B to be the proof and A to be the hypothesis. The predictor variables and features in this case are thought to be independent. That is the situation in which the existence of one characteristic would have no bearing on another. [7]

4) Decision Tree (DT)

Decision trees are supervised machine learning algorithms where each leaf node has a class label and each branch represents the result of a test on a particular parameter. At the

highest point of the tree is the parent node, as well referred to as the root node. A decision tree allows decision makers to choose the best option and move from leaf to leaf to recognize a distinct and separate class relying on the most amount of information available. Decision trees are capable of handling continuous and constant parameters. The decision tree's main benefit is that it is prone to overfitting.[9] Based on the key indicators, this algorithm divides the data into several analogous sets. The data are divided according to the entropy among each attribute, with predictors having the highest information gain or the lowest entropy:

$$E(S) = \sum_{i=1}^c -p_i \log_2 p_i \quad (4)$$

The results are simpler to read and understand. As it evaluates the set of data in the tree-like graph, this algorithm is more accurate than other algorithms. However, since only one element is checked at one time for decision-making, the data may be over classified.[10]

5) Random Forest

The tree-based classifier method used in machine learning is called random forest. The Random Forest Classifier generates multiple trees based on various attributes, and the algorithm's performance is measured by the average of the predicted outcomes of the trees. To obtain the best outcomes, it constructs several decision trees and makes use of them. For tree learning, Random Forest employs bootstrap aggregating/bagging. In a random forest, each tree contains a class expectation, and the forecast is based on the class that receives the most votes. The accuracy of the random forest classifier increases with tree count. It can handle missing values and is utilized for both regression and classification tasks, but it excels at the former. Because it requires a large amount of data and numerous trees, the findings are unpredictable in addition to dealing with taking a long time to foresee.[3], [9]

6) Support Vector Machines

SVM, also referred to as Support Vector Networks which is also a supervised learning algorithms that are used for classification and regression analysis. It uses parallel lines known as the hyperplane to divide the data points obtained by plotting in a multi - dimensional space into different categories. The maximization of the margin in between hyperplane is a requirement for data point classification. For mapping linear or nonlinear data points in a multi - dimensional space for separation, various kernels are available. Only the linear and radial basis functions were used as the kernel in our analysis.[8]

7) Artificial Neural Network (ANN)

A computational model called ANN is rooted on the composition and operations of neural networks from biology.

The purpose of the whole algorithm intends replicate its neurons found in the human brain. The artificial neural network's structure is impacted by the flow of information through the network since a neural network adapts or gets to know for each stage individually based on input and output for that stage. An ANN that is frequently can use as interconnected hidden, input and output layers that make up the Multi Layer Perceptron. Each layer is given a various number of neurons depending on the circumstances. These networks are relatively straightforward mathematical models that can improve current data analysis methods.[6]

D. Data Preprocessing

In addition to the algorithms employed, the durability of the data - set and even the preprocessing methods have an impact on the efficiency and accurateness of the predictive model. Before using machine learning techniques on a dataset, preprocessing describes the actions taken on the dataset. The dataset must be prepared and changed into a format that the machine learning techniques can understand during the data preprocessing stage.

Datasets may contain errors, incompleteness, noise, redundancies, and other issues that make them unsuitable for use directly by machine learning algorithms. The volume of data is another consideration. Some datasets have a lot of characteristics that make it challenging for the method to analyze, find patterns, or predict outcomes. By analyzing the dataset and utilizing the appropriate data preprocessing techniques, such issues can be resolved. [6]

E. Metrics for Performance Evaluation

The pre-processed data set is used in the experiment by the researchers to carry out the studies, and the explained algorithms are looked into and used. Utilizing the confusion matrix, the performance metrics listed below were used to analyze prediction models and demonstrate their performance results. The model's performance is described by the confusion matrix. The confusion matrix produced by the various ML algorithms using the proposed model.[1]

- True positive (TP): both the patient and the test are positive.
- False positive (FP): the patient does not have the disease, but the test results are positive.
- True negative (TN): the patient does not have the disease and the test results are negative.
- False negative (FN): the patient has the disease, but the test shows that it is negative.

1) Accuracy

The proportion of total correct predictions to total input samples serves as a measure of accuracy. The following formula may apply,[4]

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \quad (5)$$

2) Precision

This metric illustrates the outcome's significance and applicability. Precision, also known as positive predicted values, is the percentage of situations amongst retrieved instances that are relevant. This demonstrates how effectively the classifier handles favorable observations but says little about unfavorable ones. It is determined as,[4]

$$\text{Precision} = TP/(TP+FP) \quad (6)$$

3) Recall or Sensitivity

the relevant results are measured. The percentage of all relevant instances which being certainly retrieved is known as recall. It also goes by the name "sensitivity." By classifying it as positives, it determines the amount of true positives or model captures. It is determined as,[4]

$$\text{Recall} = TP/(TP+FN) \quad (7)$$

4) F-Measure

Combines recall and precision. The accuracy of a test is quantified by the F-measure. In order to calculate the score, the precision and recall of the experiment are both taken into account. F-score is another name for it. It is determined as,[6]

$$\text{F-Measure} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (8)$$

5) Receiver Operation Characteristic (ROC)

This is a graph that displays the performance of the classifier. Both the correctly classified instances and the incorrectly classified instances are displayed.[6]

F. Dataset Collection

The cardiovascular disease dataset from the UCI (University of California, Irvine C.A) Center for machine learning and intelligent systems is the one which is popular in research papers. It includes four hospital-related databases. There are 14 features in total across all databases, but there are various numbers of records in each. The Cleveland dataset, which has more records and fewer missing attributes than other datasets, is the one that machine learning researchers use the most. The "num" field indicates whether the patient has heart disease. It

has integer values ranging from zero (no presence) to four. There are 303 instances in the Cleveland dataset. The dataset's 14 attributes/features are listed in Table I along with a brief description of each one.[6]

TABLE I: Dataset Attributes

S. No	Attribute Name	Description
01	Age	Age in years
02	Sex	Male/female
03	Cp	Constructive pericarditis
04	Trestbps	Resting blood pressure in mmHg on admission to hospital
05	Chol	Serum cholesterol in mg/dl
06	Fbs	Fasting blood sugar (greater than 120mg/dl). Values :1=true, 0=false.
07	Restecg	Resting electrocardiographic results. Values :0=normal, 1=having ST-T wave abnormality.
08	Thalch	Maximum heart rate achieved.
09	Exang	Exercise including angina.value:1=yes,0=no
10	Oldpeak	St depression induced by exercise relative to rest
11	Slope	The slope of peak exercise ST segment.value:1=up sloping, 2=flat, 3=down sloping
12	Ca	No. of major vessels (0-3) colored by fluoroscopy
13	Thal	Inherited blood disorder that causes your body to have lesser HB than normal. Values:3=normal, 6=fixed defect, 7=reversible defect.
14	Num	Diagnosis of heart disease (angiographic disease status)

III. COMPARATIVE ANALYZE OF ML CLASSIFICATION ALGORITHMS FOR PREDICTION OF CVD

There has been ongoing research into using ML to predict heart disease for the past 20 years. The majority of papers used a variety of machine learning techniques to diagnose CVD with varying degrees of accuracy, including Decision Tree, KNN, Naive Bayes, Artificial neural network, Support vector machine, Random Forest, and logistic regression. The choice of parameters upon which techniques have been applied represents one of the points at which the papers diverge. For testing the accuracy, numerous authors have specified various methods and databases.

A. Single Approach

Many researchers use a variety of classification techniques to predict heart disease. An overview of the recent survey papers in the relevant field is provided in this section. Papers are grouped here under the Single Approach section, according to the algorithms that have been applied to their prediction models. The majority of researchers combined several algorithms or presented a contrast between them in their research work. This is covered in the section after that, which is titled as Hybrid Approach.

1) K-Nearest Neighbor (KNN)

To foresee heart diseases, Shouman M. et al. used K-Nearest Neighbor (KNN) in [11]. Through utilizing Cleveland dataset, this study demonstrates that KNN outperforms the neural network ensemble in terms of accuracy. In the paper, the outcomes of using KNN on its own and using KNN along with the voting approach were contrasts sharply. Voting is the process of breaking the data up into smaller groups and adding the classifiers towards every group. For testing, tenfold cross validation is used. The findings revealed that, depending on the value of K, the accuracy in the absence of voting varied between 94 and 97.4 percent. When the K value is 7 without any voting, the accuracy reaches its peak at 97.4%. However, unlike decision tree classification models where voting increases accuracy, voting was not able to boost the K-Nearest Neighbor accurateness for the diagnostic testing of patients with heart disease. Additionally, the findings show that when K is turned to 7 with voting, accuracy dropped to 92.7 percent.

2) Support Vector Machine (SVM)

Use of a nonlinear classification method to forecast heart diseases was recommended by R. Sharmila et al. in [12]. It emphasizes the problem of dealing large scale data for forecasting. It suggests another approach to forecasting heart problems using huge proportions of data utilizing big data tools. MapReduce and Hadoop Distributed File System, along with SVM for an optimized attribute set, are used in this study's big data tools. In order to store large amounts of data across multiple nodes and simultaneously run the SVM based prediction algorithm across various nodes, this study recommends using HDFS. Computation times are faster when SVM is used in parallel than when it is used sequentially. SVM gives greater and more efficient accuracy of 85 and 82.35 percents.

In [13], Wiharto et al. researched the diagnostic effectiveness of various SVM algorithm types using the UCI dataset. The study used a variety of SVM types, including the BTSVM, OAO, OAA, DDAG, ECOC which are stand for Binary Tree Support Vector Machine, One Against One, One Against All, Decision Direct Acyclic Graph, and Exhaustive Output Error Correction Code respectively. Firstly, the preprocessing of the data-set was performed utilizing a minimum maximum scaler. Then the algorithm was subsequently trained upon the relevant dataset using the previously stated SVM

algorithms. In this study, BTSVM outperformed the other available algorithms with an optimum accuracy of 61.86 percent in the evaluation process.

3) Naive Bayes

Naive Bayes classifier was used by Vembandasamy et al. in [14] to determine whether or not heart diseases existed. The study's data set, which included records for 500 patients with details on 11 different attributes, including the diagnosis, was obtained from one of Chennai's top centers for diabetes research. Utilizing the WEKA (Waikato Environment for Knowledge Analysis) tool, which contains a number of ML algorithms, the Naive Bayes classifier is applied. Their research's accuracy rate was 86.4198%.

Kamal Kant et al. in [15] suggested the data mining method Naive Bayes as a proposed solution for forecasting heart disease. A statistical classifier known as Nave Bayes allocates no dependence in between attributes. For identifying the class, the posterior probability must be optimised. In this situation, this classification algorithm also performs well. Nave Bayes appears to become the most successful model for disease forecasting throughout statistical probability as well as real time expert systems., guided by Decision trees and neural network.

For the purpose of predicting heart disease, Dhanashree S. Medhekar et al. in [16] proposed a classifier technique, and they also demonstrated how Nave Bayes can be applied to classification. They divided clinical expertise into five groups. No, low, average, high, and very high are the available category groups. The method will place any unidentified samples into the appropriate class label if they are found. Here, 303 records and 14 parameters from the dataset which is the coronary heart diseases research set from Cleveland Medical Institution. A coaching phase and a testing phase are the two stages of the system's operation. The classification is monitored during the coaching phase. Predictions of unknowledgeable facts or deficient values are part of the checking out segment. The accuracy of the system was 88.96%. The result demonstrates that the accuracy was attained by varying the number of occurrences within the provided dataset.

4) Decision Tree (DT)

The Decision Tree J48 algorithm was used by Sabarinathan and Sugumaran in [17] for feature selection and heart disease prediction. Thirteen medical attributes or features were included in the dataset, and 240 documents were utilized for training and 120 documents for testing. The accuracy seemed to be 75.83 percent when every features were utilised. However, accuracy climbed to 76.67 percent when feature selection was being used. Furthermore, the accuracy increased to 85% when more unimportant features were eliminated.

According to the paper, the J48 algorithm allows choosing the bare minimum of features to improve prediction accuracy.

A new model was introduced by Vikas Chaurasia et al. in [18] that improved the Decision tree's accuracy in detecting cardiovascular disease patients. Classification and Regression Tree (CART), Iterative Dichotomized 3 (ID3), and C4.5 build model are three decision tree algorithms used in this context. To increase prediction accuracy, the CART model builds a tree by repeatedly separating observations in the branches. It creates regression and classification trees to forecast categorical predictor variables and continuously dependent variables. When building the binary tree, in order to find the root first, ID3 employs an iterative inductive methodology. To describe and analyze decision situations, decision tables (DTs) use tabular representations. Data from the Cleveland Clinic Foundation are used in this study. From all 76 raw attributes, only 11 were picked. It was examined and put into use with the WEKA tool. After DT and ID3, CART seemed to have the highest accuracy which is 83.49%.

Using the WEKA tool and the UCI dataset, Jayamin Patel et al. in [19] contrasted various decision tree algorithms to determine whether or not heart disease existed. J48, logistic model tree, and random forest are various algorithms that were put to the test. Weka's C4.5 algorithms is implemented in J48, an open-source, dependable Java program. In order to further categorize the part into samples, attributes at every node are chosen in a divide-and-conquer manner as the tree is constructed. The size, which rises linearly with the examples, is, however, the biggest drawback in this situation. Through an accuracy of 56.76%, the J48 algorithm outscored the competition which has accuracy of 55.75 percent, that becomes superior to LMT algorithm.

B. Hybrid Approach

Different classification methods, including J48 that is Decision Tree, KNN, Naive Bayes, and SMO which was frequently applied for SVM training, have been assessed by Boshra Baharami et al. in [20] to determine heart disease prognosis and diagnosis. The technique of selecting dataset features based on gain ratio analysis was applied to determine the key elements. The classification algorithms are implemented using the WEKA software. The mining techniques are tested using the tenfold cross validation technique. The functionality is then analyzed and contrasted using metrics such as accuracy, precision, sensitivity, and specificity. The highest accuracy is shown by J48, which is 83.732%.

The research of S. Seema et al. in [21] focuses on methods for predicting chronic disease by mining previous medical entries for data and employing Naive Bayes, decision trees, SVM, and ANN techniques. To determine which classifiers perform better at an accurate rate, comparative research is evaluated. In this experiment, SVM has the highest accuracy, but when it comes to diabetes, Naive Bayes has the maximum accurateness.

Various algorithms, including Naive Bayes, Classification Tree, KNN, Logistic Regression, SVM, and ANN, were suggested by Ashok Kumar Dwivedi et al. in their paper [22]. In comparison to other algorithms, the Logistic Regression provides greater accuracy of 85% while NB, KNN and Classification Tree results in accuracy of 83%, 80% and 77% respectively.

On the basis of the Cleveland heart disease dataset, Pouriyeh et al. in [23] conducted an extensive contrast of separate classification algorithms to identify the classifier that outperforms all others. Radial Basis Function, Single Conjunctive Rule Learner, Naive Bayes, Multilayer Perceptron, KNN, Decision Tree, and SVM were among the classifiers used. The comparison of ensemble techniques was also included in the paper which are bagging, boosting, and stacking. The comparison of ensemble techniques was also included in the paper which are bagging, boosting, and stacking. The accuracy of classifiers was calculated by the authors using the K Fold Cross Validation methodology. Accuracy, precision, recall, F measure, and ROC curve were the performance evaluation metrics for every classifier. Different K values were tried for the KNN classifier, and when K become 9 was found to be the best value. The best combination for an artificial neural network (ANN) was determined through testing different neuron counts, and it is 13, 7, and 2 as the input, hidden, and output layers, respectively. The experiment has been composed of two experiments, including one which contrasted the various classifiers stated previously, as well as the second where the application of ensemble techniques implicated. SVM performed better than the other classification methods in the first experiment, with an accuracy of 84.15 percent, according to the results. The boosting technique using for SVM also claimed as being the greatest effectiveness, including an accuracy of 84.81 percent, throughout the second experiment.

A productive intelligent system for supporting medical decisions was created by Zriqat et al. in [24]. Five classification algorithms which are Naive Bayes, Support Vector Machine, Decision Tree, Random Forest and Discriminant were compared. On the Statlog Heart Disease and the Cleveland Heart Disease datasets, the analysis was carried out using MATLAB. According to the results, Decision Tree performed with the maximum accuracy scoring 99.01 percent for the Cleveland dataset and 98.15 percent for the Statelogs dataset for each datasets.

In order to evaluate heart disease prognostication, G Purusothaman et al. in [25] reviewed and compared various classification techniques. The authors highlight the usage of hybrid models, that further combine multiple classification algorithms, rather than using a single methodology including the decision tree, ANN, or Naive Bayes. They have studied the research done by scientists who looked into the performance of hybrid models. Single models like Decision Tree, ANN, and Naive Bayes perform 76, 85, and 69 percents better than average, respectively. Also the accuracy rate of hybrid approaches, however, is 96%. As a conclusion, hybrid models produce trustworthy and promising classifiers that can accurately predict heart diseases.

A model for predicting heart disease was created by Marjia et al. in [26] using KStar, J48, SMO, Bayes Net, and Multilayer perception with WEKA software. SMO and Bayes Net perform better than KStar, Multilayer Perception, and J48 techniques employing K fold cross validation when considering a variety of performance metrics. KStar, J48, SMO, Bayes Net and Multilayer Perceptron shows accuracy of 75%, 86%, 89%, 87%, 86% respectively. These algorithms' accuracy performances are yet insufficient. As a result, the performance of accuracy is further enhanced to deliver greater disease diagnosis decisions.

In [27], Khateeb and Usman tested a number of classification algorithms using UCI Cleveland dataset, including Naive Bayes classifier, KNN, decision trees, and bagging technique. Six cases were used to divide the work, and each classifier determined the accuracy for each case. First Case involved the dataset being subjected to the full set of classifiers without any feature reduction. Second Case involved the use of feature reduction, in which only seven of the dataset's 14 attributes that the ones most crucial for the diagnosis of heart disease were chosen for use. Third Case only had its most standardized characteristics such as age, sex, and resting blood sugar were removed.

A model for forecasting heart disease was developed by Dangare and Apte in [28]. The dataset is composed of the Statlog database, which has 270 records, and the Cleveland database, which has 303 records. They introduced 2 more attributes which are obesity and smoking, in addition to the thirteen existing attributes in the dataset. The dataset was preprocessed using the WEKA tool. Decision Tree, Naive Bayes, and ANN were the classification methods used to analyze the dataset. The accuracy of ANN was 100%, that of Decision Tree was 99.62%, and that of Naive Bayes was 90.74 %, highlighting that ANN is the best algorithm.

The Heart Disease Prediction System using Data Mining Approaches was proposed by MeghaShahi et al. in [29]. In healthcare facilities, WEKA software is used to provide quality service and make automatic diagnoses of disease. The paper made use of a number of algorithms, including Decision Tree, Naive Bayes, Association rule, KNN, and Naive Bayes. This study came up with the conclusion that SVM is more accurate and efficient than other data mining techniques.

In [30], Sharan Monica.L et al. recommended conducting a research on cardiovascular disease prediction. In this study, data mining methods were suggested as a way to forecast the disease for the benefit of healthcare professionals, it aims to provide an overview of the most recent information of current techniques from dataset extracting. On the basis of how long it took to construct the system's decision tree, one can determine performance. The main objective is to forecast the heart disease using the fewest possible attributes. Models with forecast abilities have been created using naive Bayes classifier, ideally for continuous datasets. Significant data relationships are quickly displayed using CART. Application of these 3 algorithms was done through WEKA. While CART provided the maximum accuracy of 92.2 percent, Naive Bayes and J48 provided 88.5% and 91.4% accuracy respectively. J48 used to be the fastest to construct in 0.08 seconds.

TABLE II: Comparative Analyze of ML Classification Algorithms for Prediction of Heart Diseases

<i>Author</i>	<i>Classification Techniques</i>	<i>Tool</i>	<i>Dataset Used</i>	<i>Best Technique</i>	<i>Accuracy</i>
Shouman M. et al.[11]	KNN	Not Reported	Cleveland (UCI)	Not Applicable.	97.4%
R. Sharmila et al.[12]	SVM	HDFS, Mapreduce	Cleveland (UCI)	Not Applicable.	<i>SVM in parallel fashion is more accurate than sequential SVM</i>
Wiharto et al.[13]	SVM	Not Reported	Cleveland (UCI)	BTSVM	61.86%
Vembandasa my et al. [14]	NB	WEKA	A diabetic research institute in Chennai	Not Applicable.	86.4198%
Kamal Kant et al. [15]	NB	Not Reported	Not mentioned	Not Applicable.	<i>NB, followed by Neural Networks.</i>
Dhanashree S. Medhekar et al. [16]	NB	Not Reported	Cleveland (UCI)	Not Applicable.	89.58%
Sabarinathan and Sugumaran [17]	DT	Not Reported	A dataset with 240 records for testing and 120 for training	J48 with feature selection	85%
Vikas Chaurasia et al. [18]	DT	WEKA	Cleveland (UCI)	CART	83.49%
Jayamin Patel et al. [19]	DT	WEKA	Cleveland (UCI)	J48	56.76%
Boshra Baharami et al. [20]	J48, NB, KNN, SMO	WEKA	Dataset contains 209 records and 8 features that is collected from a hospital in Iran	J48	83.732%.

					Highest Accuracy
S. Seema et al. [21]	NB, DT, SVM	Not Reported	Cleveland (UCI)	SVM in case of heart disease And NB in case of diabetes	Case of heart disease SVM: 95.556% Case of diabetes NB: 73.588%
Ashok Kumar Dwivedi et al. [22]	NB, KNN, Logistic Regression, Classification Tree	Not Reported	Statlog (UCI)	Logistic Regression	85%
Pouriyeh et al. [23]	NB, DT, MLP, KNN, SCRL, RBF, SVM, bagging, boosting and stacking	Not Reported	Cleveland (UCI)	Boosting with SVM	84.81%
Zriqat et al. [24]	NB, DT, Discriminant, Random Forest, and SVM	MATLAB	Cleveland and Statlog (UCI)	DT	99.01% for Cleveland and 98.15% for Statlog
G Purusothaman et al. [25]	DT, NB, ANN	Not Reported	Not mentioned	Hybrid Approach	96%
Marjia et al. [26]	K Star, J48, SMO, Bayes Net, Multilayer Perception	WEKA	Cleveland (UCI)	SMO	89%
Khateeb and Usman [27]	NB, KNN, DT and bagging technique	WEKA	Cleveland (UCI)	KNN	79.20%
Dangare and Apte [28]	DT, NB and ANN	WEKA	Cleveland and Statlog (UCI)	ANN	Almost 100%
MeghaShahi et al. [29]	SVM, NB, Association rule, KNN, ANN, and DT	WEKA	Not mentioned	SVM	SVM is effective and provides more accuracy
Sharan Monica.L et al. [30]	J48, NB, Simple CART	WEKA	Not mentioned	Simple CART	92.2%

IV. DISCUSSION

This article outlines some recent machine learning research on cardiovascular diseases. Numerous papers that used techniques for machine learning classification were analyzed and structured. The tools, datasets, and number of proposed models used all have an impact on the proposed models' accuracy. The massive amounts of data produced by the healthcare sector can be effectively extracted for relevant info using machine learning algorithms. These studies demonstrate that using a combination of ML techniques produces significantly better results than using just one ML technique upon the data set. For the majority of research projects, Java is the programming language of choice. Among the other widely used tools for data analysis are WEKA, Matlab, etc.

A system for managing heart disease can be quickly and efficiently implemented with careful choice of the ML techniques and accurate application of such techniques on given dataset. The necessary dataset is split into two, one of set which is used for classification and the other, which is smaller, is used for verification. The tenfold cross validation method is frequently employed. Among the works some were compare various classification algorithms on a dataset to determine whether or not an assigned patient has any chance of having a heart condition. Other papers have functioned on extracting from a dataset the factors that contribute to heart diseases.

Heart disease prognosis using machine learning is a crucial field that can benefit patients as well as medical experts. Although there is a ton of patient data available in medical centers, not most of this can be published because the field is yet expanding. As can be seen in table two, the UCI repository is where the majority of studies obtained their data sources. Although this dataset's quality plays a crucial role in how accurately a prediction is made, further health professionals must be inspired to publish high quality sets of data while maintaining patient privacy, as then research teams have a reliable source which allow them to build their models and get successful outcomes.

The dataset's attributes and records, the preprocessing methods used, and also the classifier used to construct the model. It varies based on if the model is a hybrid one as well as if it is employed feature selection. According to table 2, which mention above Dangare and Apte have applied an Artificial Neural Network using the WEKA tool and as a result, a mixture of the Cleveland and Statlog heart disease datasets produces the research with the highest accuracy which is nearly 100%.

Decision tree, Naive Bayes, ANN, KNN, and SVM are mostly used common classification techniques. It finally appears that Decision Tree and Artificial Neural Network operated similarly and more accurately in the majority of models for forecasting heart diseases. Some recent works have evaluated

hybrid models in addition to analyzing these widely utilized methods. To achieve better results, a hybrid model combines several mostly known classification and selection approaches into a single model. If the right blends of various algorithms are selected, hybrid models are seen to provide very high accuracy.

V. CONCLUSION

Left untreated heart conditions can become unmanageably severe. Heart diseases have always been complex and claim many lives each year. The patient may suffer severe consequences either in a brief period if such early signs of heart problems are neglected. The problem has gotten worse as a result of modern society's unhealthy lifestyles and high levels of stress. Heart disease and stroke risk factors include using cigarettes and skipping meals. Earlier stage disease detection allows for control of the condition. But it's always a good idea to work out every day and get rid of harmful habits as soon as possible.

This study looks at various analyzed systems for predicting heart disease using various classification methods in an effort to predict the condition more precisely and effectively. According to the study, each paper's technique uses a separate count of attributes and classifiers. Every paper has been found to have a unique accuracy to it. However, it is necessary to have a trustworthy model for determining whether heart disease will develop on given known and unknown threat elements. Death can occasionally result from weak clinical judgments. Since time is the essence when it comes to heart diseases, many patients can be saved by accurate identification of risks at the appropriate moment.

In summary, to forecast the presence of heart disease, numerous machine learning approaches have been constructed. Discover how well each algorithm predicts outcomes, then add the recommended system into place where it is required. To increase the accuracy of algorithms, employ more appropriate feature classification techniques. There were also many treatment options available for patients who were diagnosed with a particular type of heart disease. From such a preferable dataset, data mining can generate knowledgeable information. The database will gain knowledge more as more when data is added, which will make a system that is more intelligent.

ACKNOWLEDGMENT

With all of the progress of this study up to this point, Dr. (Mrs.) Fernando K.S.D.'s assistance was a tremendous source of strength and inspiration for me to accomplish this paper and encourage me with directing towards the completing this writing. I would like to pay my sincere gratitude for the

guidance and valuable suggestions which inspired me on this accomplishment.

REFERENCES

- [1] A. Rajdhan and A. Agarwal, 'Heart Disease Prediction using Machine Learning', *Int. J. Eng. Res.*, vol. 9, no. 04, p. 4.
- [2] S. Islam, N. Jahan, and Mst. E. Khatun, 'Cardiovascular Disease Forecast using Machine Learning Paradigms', in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, Mar. 2020, pp. 487–490. doi: 10.1109/ICCMC48092.2020.ICCMC-00091.
- [3] M. Furqan, H. Rajput, S. Narejo, A. Ashraf, and K. Awan, 'Heart Disease Prediction using Machine Learning Algorithms', p. 6.
- [4] A. Agrahary, 'Heart Disease Prediction Using Machine Learning Algorithms', *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, pp. 137–149, Jul. 2020, doi: 10.32628/CSEIT206421.
- [5] M. Tech, Scholar, Department of Computer Science Engineering, All Saint College of Technology, Bhopal (MP), India., P. Sharma, S. Site, and Department of Computer Science Engineering, All Saint College of Technology, Bhopal (MP), India., 'A Comprehensive Study on Different Machine Learning Techniques to Predict Heart Disease', *Indian J. Artif. Intell. Neural Netw.*, vol. 2, no. 3, pp. 1–7, Apr. 2022, doi: 10.54105/ijainn.C1046.042322.
- [6] M. I. Al-Janabi, M. H. Qutqut, and M. Hijawi, 'Machine Learning Classification Techniques for Heart Disease Prediction: A Review', *Int. J. Eng.*, p. 8.
- [7] M. N. R. Chowdhury, E. Ahmed, Md. A. D. Siddik, and A. U. Zaman, 'Heart Disease Prognosis Using Machine Learning Classification Techniques', in *2021 6th International Conference for Convergence in Technology (I2CT)*, Maharashtra, India, Apr. 2021, pp. 1–6. doi: 10.1109/I2CT51068.2021.9418181.
- [8] P. S. Kohli and S. Arora, 'Application of Machine Learning in Disease Prediction', in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India, Dec. 2018, pp. 1–4. doi: 10.1109/CCAA.2018.8777449.
- [9] N. Nissa, S. Jamwal, and S. Mohammad, 'Heart Disease Prediction using Machine Learning Techniques', vol. 13, no. 67, p. 13.
- [10] P. Anbuselvan, 'Heart Disease Prediction using Machine Learning Techniques', *Int. J. Eng. Res.*, vol. 9, no. 11, p. 4.
- [11] M. Shouman, T. Turner, and R. Stocker, 'Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients', *Int. J. Inf. Educ. Technol.*, pp. 220–223, 2012, doi: 10.7763/IJNET.2012.V2.114.
- [12] R. Sharmila and S. Chellammal, 'A conceptual method to enhance the prediction of heart diseases using big data Techniques', *Int. J. Comput. Sci. Eng.*, p. 5, 2018.
- [13] Wiharto, H. K. usnanto, and Herianto, 'Performance Analysis of Multiclass Support Vector Machine Classification for Diagnosis of Coronary Heart Diseases', *Int. J. Comput. Sci. Appl.*, vol. 5, no. 5, pp. 27–37, Oct. 2015, doi: 10.5121/ijcsa.2015.5503.
- [14] K. Vembandasamy, R. Sasipriya, and E. Deepa, 'Heart Diseases Detection Using Naive Bayes Algorithm', vol. 2, no. 9, p. 4.
- [15] K. Kant, 'IJSRD - International Journal for Scientific Research & Development| Vol. 2, Issue 04, 2014 | ISSN (online): 2321-0613', vol. 2, no. 04, p. 3.
- [16] Dhanashree S. Medhekar, '[No title found]', *Int. J. Enhanc. Res. Sci. Technol. Eng.*.
- [17] V. Sabarinathan and V. Sugumaran, 'Diagnosis of Heart Disease Using Decision Tree', vol. 2, no. 6, p. 7, 2014.
- [18] V. Chaurasia and S. Pal, 'Early Prediction of Heart Diseases Using Data Mining Techniques', p. 11, 2013.
- [19] J. Patel and D. S. Patel, 'Heart Disease Prediction Using Machine learning and Data Mining Technique', p. 9.
- [20] B. Bahrami and M. H. Shirvani, 'Prediction and Diagnosis of Heart Disease by Data Mining Techniques', vol. 2, no. 2, p. 5, 2015.
- [21] K. Deepika and S. Seema, 'Predictive analytics to prevent and control chronic diseases', in *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, Bangalore, India, 2016, pp. 381–386. doi: 10.1109/ICATCCT.2016.7912028.
- [22] A. K. Dwivedi, 'Performance evaluation of different machine learning techniques for prediction of heart disease', *Neural Comput. Appl.*, vol. 29, no. 10, pp. 685–693, May 2018, doi: 10.1007/s00521-016-2604-1.
- [23] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, 'A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease', in *2017 IEEE Symposium on Computers and Communications (ISCC)*, Heraklion, Greece, Jul. 2017, pp. 204–207. doi: 10.1109/ISCC.2017.8024530.
- [24] I. A. Zriqat, A. M. Altamimi, and M. Azzeh, 'A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods', vol. 14, no. 12, p. 13, 2016.
- [25] G. Purusothaman and P. Krishnakumari, 'A Survey of Data Mining Techniques on Risk Prediction: Heart Disease', *Indian J. Sci. Technol.*, vol. 8, no. 12, Jun. 2015, doi: 10.17485/ijst/2015/v8i12/58385.
- [26] M. Sultana, A. Haider, and M. S. Uddin, 'Analysis of data mining techniques for heart disease prediction', in *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, Dhaka, Bangladesh, Sep. 2016, pp. 1–5. doi: 10.1109/CEEICT.2016.7873142.
- [27] N. Khateeb and M. Usman, 'Efficient Heart Disease Prediction System using K-Nearest Neighbor Classification Technique', in *Proceedings of the International Conference on Big Data and Internet of Thing - BDIOT2017*, London, United Kingdom, 2017, pp. 21–26. doi: 10.1145/3175684.3175703.
- [28] C. S. Dangare and S. S. Apte, 'Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques', *Int. J. Comput. Appl.*, vol. 47, no. 10, pp. 44–48, Jun. 2012, doi: 10.5120/7228-0076.
- [29] M. Shahi and R. K. Gurm, 'Heart Disease Prediction System using Data Mining Techniques', *Orient J. Comput. Sci. Technol.*, vol. 6, 2017.
- [30] S. Monica.L and S. Kumar.B, 'Analysis of CardioVascular Disease Prediction using Data Mining Techniques', *Int. J. Mod. Comput. Sci.*, vol. 4, Feb. 2016.