DOI: 10.24996/ijs.2022.63.9.28

Iraqi Journal of Science

ISSN: 0067-2904

Heart Disease Classification–Based on the Best Machine Learning Model

Melad Mizher Rahma [1]*, Aymen Dawood Salman [2]

[1] *Department of Computer Science, Information Institute for Higher Studies, Iraqi Computer Informatics Authority, Baghdad, Iraq*

[2] *Department of Computer Engineering, University of Technology, Baghdad, Iraq*

Received: 19/9/2021 Accepted: 21/11/2021 Published: 30/9/2022

Abstract

In recent years, predicting heart disease has become one of the most demanding tasks in medicine. In modern times, one person dies from heart disease every minute. Within the field of healthcare, data science is critical for analyzing large amounts of data. Because predicting heart disease is such a difficult task, it is necessary to automate the process in order to prevent the dangers connected with it and to assist health professionals in accurately and rapidly diagnosing heart disease. In this article, an efficient machine learning-based diagnosis system has been developed for the diagnosis of heart disease. The system is designed using machine learning classifiers such as Support Vector Machine (SVM), Nave Bayes (NB), and K-Nearest Neighbor (KNN). The proposed work depends on the UCI database from the University of California, Irvine for the diagnosis of heart diseases. This dataset is preprocessed before running the machine learning model to get better accuracy in the classification of heart diseases. Furthermore, a 5-fold cross-validation operator was employed to avoid identical values being selected throughout the model learning and testing phase. The experimental results show that the Naive Bayes algorithm has achieved the highest accuracy of 97% compared to other ML algorithms implemented.

Keywords- Machine Learning, Heart Disease (HD), Naïve Bayes (NB) , KNN , SVM;

تصنيف أمراض القلب – استنادًا إلى أفضل نموذج للتعلم الآلي

ميلاد مزهر رحمة [1]*, أيمن داود سلمان [2]

[1] قسم علوم الحاسوب ، معهد المعلومات للدراسات العليا ، الهيئة العراقية لمعلوماتية الحاسوب ، العراق

[2] أيمن داود سلمان، قسم هندسة الحاسوب ،الجامعة التكنولوجية، العراق

الخلاصة

في السنوات الأخيرة ، أصبح التنبؤ بأمراض القلب أحد أكثر المواقف صعوبة في الطب. في العصر الحديث يموت شخص واحد من أمراض القلب كل دقيقة. في مجال الرعاية الصحية ، يعد علم البيانات أمرًا

* Email: Ms201930544@ipps.icci.edu.iq

بالغ الأهمية لتحليل كميات كبيرة من البيانات. نظرًا لأن التنبؤ بأمراض القلب مهمة صعبة ، فمن الضروري أتمتة العملية من أجل منع المخاطر المرتبطة بها ومساعدة المهنيين الصحيين في التشخيص الدقيق والسريع لأمراض القلب. في هذا البحث، تم تطوير نظام تشخيص فعال قائم على التعلم الآلي لتشخيص أمراض القلب. تتضمن مصنفات التعلم الآلي استخدام Support Vector Machine و Naïve Bayes و KNN في تصميم النظام. يعتمد النظام المقترح على قاعدة بيانات UCI من جامعة كاليفورنيا في إيرفين لتشخيص أمراض القلب. تتم معالجة مجموعة البيانات هذه مسبقًا قبل تشغيل نموذج التعلم الآلي للحصول على دقة أفضل في تصنيف أمراض القلب. علاوة على ذلك ، تم استعمال عامل التحقق من صحة 5 أضعاف لتجنب اختيار القيم المتطابقة خلال مرحلة التعلم والاختبار النموذجي. أظهرت النتائج التجريبية أن خوارزمية Naive Bayes حققت أعلى دقة بلغت 97٪ مقارنة بخوارزميات ML الأخرى المطبقة.

I. Introduction

Heart disease (HD) is typically considered to be one of the most complicated and life-threatening illnesses in humans. As a result of this disease, the heart is generally not able to pump the specified quantity of blood to different elements of the body to carry out the body’s ordinary activities. As a result, cardiac failure arises [1]. Within the United States, the rate of heart sickness is quite high [2]. The signs and symptoms of HD encompass chest pain, swollen feet, weakness of the physical body, and fatigue with associated signs such as elevated jugular venous pressure and peripheral edema, which may be produced by functional cardiac or no cardiac abnormalities [3]. The early detection methods for HD were challenging, and the resulting uncertainty became one of the principal problems impacting people's quality of life [4]. HD diagnosis and remedy are distinctly challenging, in particular in poor countries, because of the loss of clinical equipment, physicians, and different services, all of which have an effect on the right diagnosis and remedy of heart patients [5]. Correct and accurate identification of a patient's heart attack risk is critical for reducing the risk of significant heart problems and enhancing heart protection [6]. According to the European Society of Cardiology, 26 million people worldwide have been diagnosed with HD. Last year, 3.6 million new cases were diagnosed. Half of patients with HD die within two years, and heart disease management costs account for around 3% of healthcare expenditure [7]. An invasive HD diagnosis is based on a review of the patient's health history, a clinical assessment report, and a medical expert's examination of the patient's symptoms. Due to human error, many of these techniques result in incorrect diagnoses and, in many cases, delays in diagnosis outcomes. It is also more costly and computationally complicated, and it takes longer to determine [8]. To overcome the challenges of invasive-based heart disease diagnosis, ML prediction models-based noninvasive medical decision support systems such as K-NN, SVM, NB, DT, ANN, Logistic Regression (LR), Ada-Boost (AB), FL, and rough set theory have been developed by various researchers and are commonly utilized for HD diagnosis [9 and 10]. As a consequence of these ML-based expert medical decision systems, the ratio of HD fatalities has been reduced [9]. Several research projects have focused on utilizing a machine-learning-based approach to diagnose HD, for example. The classification success of several ML methods on the Cleveland HD dataset was recorded in a literature review [10], and [11]. Many researchers have utilized this dataset to study various classification challenges linked to heart diseases using various machine learning classification techniques. For example, predict coronary artery disease in its early stage so that patients can undergo treatment and save their lives [12]. The main contribution of this work is to find the best classifier for the classification of heart diseases and to help physicians diagnose the heart condition of their patients with the highest degree of accuracy and efficiency.

II. RELATED WORKS

Alotaibi [13] has developed a machine learning model that compares five distinct approaches. The Rapid Miner tool outperformed MATLAB and Weka in terms of accuracy. This study looked into the accuracy of the classification approaches: Decision Tree, Logistic Regression, Random Forest, Naive Bayes, and SVM. The decision tree algorithm was found to be the most accurate. Latha et al. [14] performed a comparative analysis to improve the predictive accuracy of heart disease risk using ensemble techniques on the Cleveland dataset of 303 observations. They applied the brute force method to obtain all possible attribute set combinations and trained the classifiers. They achieved a maximum increase in the accuracy of a weak classifier of 7.26% based on the ensemble algorithm and produced an accuracy of 85.48% using a majority vote with NB, BN, RF, and MLP classifiers using an attribute set of nine attributes. Mohan et al. [15] developed an effective hybrid random forest with a linear model (HRFLM) to enhance the accuracy of heart disease prediction using the Cleveland dataset with 297 records and 13 features. They concluded that the RF and LM methods provided the best error rates. Louridi et al. [16] proposed a solution to identify the presence/absence of heart disease by replacing missing values with the mean values during preprocessing. They trained three machine learning algorithms, namely, NB, SVM (linear and radial basis function), and KNN, by splitting the Cleveland dataset of 303 instances and 13 attributes into 50:50, 70:30, 75:25, and 80:20 training and testing ratios. Gupta et al. [17] replaced the missing values based on the majority label and derived 28 features using the Pearson correlation coefficient from the Cleveland dataset and trained LR, KNN, SVM, DT, and RF classifiers using the factor analysis of mixed data (FAMD) method; the results based on a weight matrix RF achieved the best accuracy of 93.44%. Perumal et al. [18] developed a heart disease prediction model using the Cleveland dataset of 303 data instances through feature standardization and feature reduction using PCA, where they identified and utilized seven principal components to train the ML classifiers. They concluded that LR and SVM provided almost similar accuracy values (87% and 85%, respectively) compared to that of k-NN (69%). Kumar et al. [19] trained five machine learning classifiers, namely, LR, SVM, DT, RF, and KNN, using a UCI dataset with 303 records and 10 attributes to predict cardiovascular disease. The RF classifier achieved the highest accuracy of 85.71% with a ROC AUC of 0.8675 compared to the other classifiers. Gazeloglu et al. [20] projected 18 machine learning models and 3 feature selection techniques (correlation-based FS, chi-square, and fuzzy rough set) to find the best prediction combination for heart disease diagnosis using the Cleveland dataset of 303 instances and 13 variables. Sharma et al. [21] used the heart disease dataset, which is available in the (UCI) machine learning repository and was employed in their research. Using data mining strategies such as NB, DT, LR, and RF, the suggested system predicts the likelihood of HD and classifies patient risk levels. As a result, in their work, they have been capable of evaluating the output of numerous ML algorithms. The outcomes show that the RF method has the very best accuracy of 90.16% when compared to different ML techniques. Pavithra et al. [22] proposed a new hybrid feature selection technique with the combination of random forest, AdaBoost, and linear correlation (HRFLC) using the UCI dataset of 280 instances to predict heart disease. Eleven (11) features were selected using filter, wrapper, and embedded methods; an improvement of 2% was found for the accuracy of the hybrid model. Kavitha et al. [23] implemented a novel hybrid model on the Cleveland heart dataset of 303 instances and 14 features with a 70:30 ratio for training and testing by applying DT, RF, and hybrid (DT + RF) algorithms.

III. THEORETICAL BACKGROUND

A- HEART DISEASE

The human heart is a crucial organ that serves as a pump to circulate blood throughout the body. If the body's blood circulation is poor, organs such as the brain suffer, and if the heart entirely stops pumping, within minutes, death happens. The heart's correct functioning is critical for survival. Problems with the heart and its blood vessels are referred to as "heart disease".

1. The common form of cardiac disease worldwide is a condition in which plaque builds up in the arteries of the heart, causing the heart to receive less blood and oxygen.
2. Angina pectoris is a condition in which a person suffers from chest pain. It is a medical term for chest discomfort that occurs as a result of a loss of blood flow to the heart. It is also called angina, and it is an indication that you are having a heart attack. The chest discomfort comes in waves that last a few seconds or minutes.
3. Cardiomyopathy is the weakening or change of the heart muscle's structure as a result of inadequate cardiac pumping. Hypertension, alcohol use, viral infections, and genetic abnormalities are all frequent causes of cardiomyopathy.
4. Arrhythmias It is linked to a problem with the heartbeat's rhythmic action. The heartbeat might be irregular, slow, or fast. These irregular heartbeats are caused by a short circuit in the electrical circuitry of the heart.
5. Myocarditis is an inflammation of the heart muscle caused by viral, fungal, or bacterial diseases. It's an uncommon disease with minimal symptoms such as joint pain, leg swelling, or fever that isn't caused by the heart [24], [25].

B. Naïve Bayes

The Naive Bayes predictor was utilized in this research. It is also a supervised learning approach to categorize data by calculating the likelihood of independent factors. The high likelihood class is allocated to the whole transaction after the probability of each class is calculated [26], [27]. In different datasets, such as instructional data mining [29] and health data mining, NB is a popular approach for predicting classes [30]. This model may be used to categorize a variety of datasets, such as sentiment analysis [30] and virus identification [25]. It operates by predicting a predefined class for each document based on the values of independent variables. It calculates the likelihood of A given that B, as illustrated in the equation below [29]. Then focus on identifying a distinct class for each feature; in this situation, none of the other variables are interdependent [29]. The probability is calculated using the below equation:

$$p(C|X) = \frac{p(X|C)\, p(c)}{p(X)} \qquad (1)$$

P (C|X): Posterior probability of class (c, target) given predictor (x, attributes).

P (C): Prior probability of class.

P (X|C): Likelihood or posterior probability of X conditioned on Ci.

P (X): Prior probability of X.

C. K Nearest Neighbor Algorithm

The KNN method is an example-based learning algorithm that is frequently utilized in real-life scenarios. Both classification and regression problems may be solved using the KNN method. The K-Nearest Neighbor technique is another name for lazy learning. In comparison to other classification approaches, it is a simple classification method with a low computing cost. In a j-dimensional dataset, for each sample analyzed, the K closest sample is found. Common distance computation methods are used to determine the distance between the samples and the sample being assessed, like Euclid, Hamming, and Manhattan. The sample's class can be decided by a majority vote of the K closest sample classes [25]. The following equation can be given for the Euclid distance calculation method:

$$\text{d}= \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \qquad (2)$$

Where p stands for the sample evaluated, q stands for any sample within the training dataset, and n stands for the feature size**.**

D. SVM Algorithm

The SVM is the final ML algorithm used in this research. This is also known as a supervised (ML) model, because the classes in the database are predefined [31]. It operates by classifying the items in the collection into specified categories. In order to improve accuracy, it classifies transactions by assigning one or more classes [32]. Previously, SVM has already been used in a medical data application to predict the correct class for HD patients [33].

E. Data Overview

The (HD) dataset, which was obtained from the Kaggle platform, was used in this study [34]. The data came from four different databases in total, but just the Cleveland data was used in this study. It is an open dataset with many properties, but for this experiment, just 14 were chosen, as stated and recommended by several researchers who believe that the selected 14 attributes are the most effective in predicting heart disease in a patient [27]. A total of 303 patients' records are also included in the database file. Table 1 shows the full explanation of each property as well as the number of possible values.

Table 1-Data -Overview and Properties Description

S.No	Attribute Description	Differentiated Values
1	**Age -** The person's age is the first feature to define [Min : 29, Max : 77]	range of values between 29 and 77
2	**Sex** - The gender of a patient is the second attribute.[“0”denotes female, whereas “1” denotes male]	0, 1
3	**CP -**The third characteristic describes the level of a patient's chest pain (CP) when they arrive at the hospital. This characteristic has four different types of values, each of which describes a different amount of chest discomfort	0, 1, 2, 3
4	**RestBP -** This attribute described patient's blood pressure (BP) while in the hospital [Minimum blood pressure: 94, maximum blood pressure: 200]	Multiple values between 94 and 200
5	**(Chol)-** The cholesterol level is displayed in this column [Cholesterol Minimum: 126, Cholesterol Maximum: 564].	Multiple values between 126 and 564
6	**(FBS)-** The patient' fasting blood sugar level is described in the next attribute. It has values that are classified as binary. If the patient has more than 120mg/dl sugar, the result is 1, else it is 0.	0,1
7	**(RestECG)-**This parameter shows the ECG result on a scale of 0 to 2. Each number indicates the degree of the pain.	0, 1, 2
8	**(HeartBeat)-** The maximum value of heartbeat recorded at the time of admission [Minimum: 71, Maximum: 202]	between 71 and 202

9	**Exang -** This measure was used to determine whether or not exercise causes angina. If yes, the value will be “1,” and if not, it will be “0.”	0, 1
10	**oldPeak-** The patient's depressive status is the next property to define.	There are a variety of real number values between 0 and 6.2
11	**Slope-** The patient’ state through peak exercise.	1, 2, 3
12	**CA-** This property displays the fluoroscopy status. It depicts the number of colored vessels.	0, 1, 2, 3
13	**Thal-** This is a different type of test that is necessary when a patient has chest discomfort or trouble breathing. The outcome of the Thallium test is represented by four different types of values.	0, 1, 2, 3
14	**Target –**This column is often referred to as the Class or Label column.	0,1

IV. PROPOSED METHODOLOGY

The proposed work predicts HD through exploring the three abovementioned algorithm types as well as the overall performance analysis The goal of this work is to efficiently classify five cases of the heart (arrhythmia, myocardial infarction, ischemic, high blood cholesterol, and normal cases of the heart). The entered values from the patient's clinical report are entered by the health professional. The information is fed into a model that estimates the likelihood of developing HD. Figure 1 illustrates the whole operation involved.

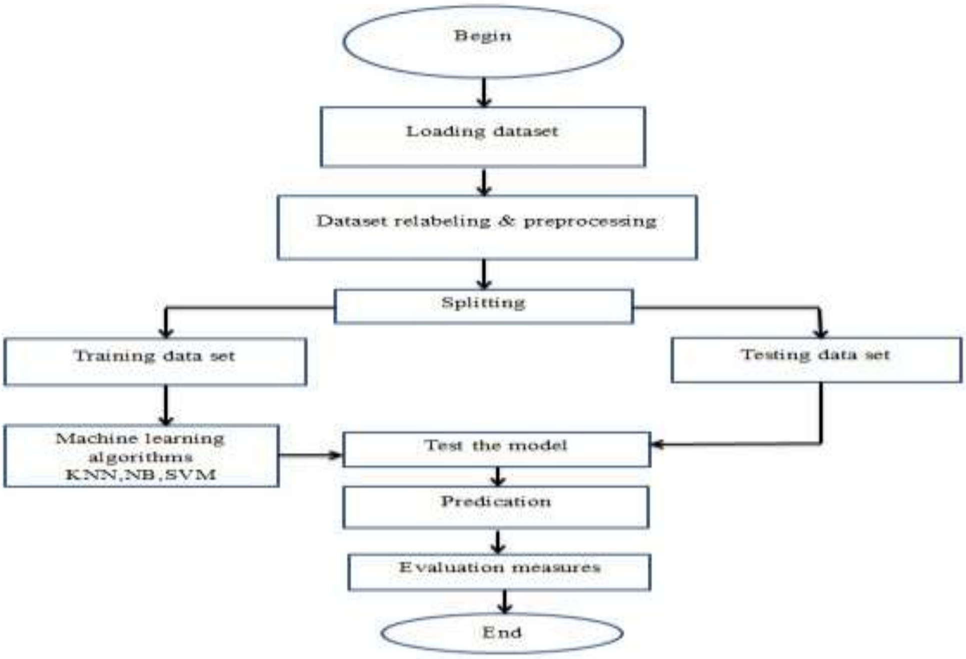

Figure 1-Flowchart illustrated the design of proposal classifier

A-Dataset Relabeling and Preprocessing

This study's dataset was obtained from the UCI Machine Learning Repository [34], which is a dataset of the open-source type having a number of characteristics, but for this experiment, just fourteen were chosen, as explained and proposed by many researchers. The researchers believed that the selected 14 attributes were the most effective in predicting heart disease in a patient. This dataset was relabeled by the written-algorithm using the Pandas library in Python that works on relabeling data to five categories (0-4) according to medical consultations and international research that were studied on the most important symptoms and important tests in diagnosing heart diseases. Table 2 illustrates the algorithm that is used to do relabeling on a dataset.

Table 2-Illustrate the algorithm that is used to do labeling on a dataset.

```
Algorithm shows relabeling dataset for five classes

Input: dataset of heart diseases

Output: relabel dataset from 2 classes to 5 classes

If(ECG=1 & chest pain type=1,3 & Thalah(max heart rate): heart rate>100 & Exang: exercise induced angina=1 & st.Slope=1 &Chol(cholesterol):cholesterol>200)
=>   Myocardil Infarction=1

else If(cp: chest pain type=2 &Thalah(max HR):heart rate>100 & Exang: exercise induced angina=1 & Chol(cholesterol): cholesterol >200
=> Ischemic=2

else If(Chol(cholesterol):cholesterol >200)
=>Cholestrol=3

else If(Thalah(max heart rate): heart rate>100)
=>Arrhythmia=4

else =>healthy=0
```

After the data relabeling process, the dataset is preprocessed, which includes:

1-In order to prepare data for analysis, **data cleaning** is the act of eliminating or modifying data that is incorrect, incomplete, irrelevant, redundant, or poorly organized [35].

2-Data transformation, by using the Discretization process to convert continuous data into a set of data intervals [35].

B-Data Splitting

After the data has been prepared through preprocessing, the splitting process is performed to divide it. The data set is divided into two parts: a (training) data set and a (testing) or a (validation) data set, with the "training" data set being used to train the model and the "test" or "validation" data set being used to qualify performance. In this work, 50% was determined as test data and 50% as training data.

C- Classification

This research used three supervised learning classification models: Naïve Bayes , KNN, and SVM to classify five cases of heart disease. A 5-fold cross-validation operator was employed to avoid identical values being selected throughout the model learning and testing phase. It assists in dividing the data into k equal groups and allows each subset to participate in the training and testing phases. The cross-validation operator is thought to be efficient since it repeats the learning phase k times, with each testing data choice differing from the previous. Finally, the experiment is repeated k times and the average findings are utilized. For learning and testing purposes, cross-validation is a widely used operator. It affords the data choice in 4 exclusive ways: shuffled sampling, liner sampling, stratified sampling, and automatic [36]. In this study, however, shuffled sampling is used.

V. COMPARISONS AND DISCUSSION OF MODEL PERFORMANCES

In terms of accuracy, PPV, NPV, recall, specificity, and F-measure, the heart disease classification model based on the Nave Bayes classifier is compared to the current KNN and SVM, as shown in table 4. The primary parameters which might be evaluated are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values.

The assessment criteria are shown below:

A-Accuracy (Acc): is a statistical measure of a classifier's ability to properly identify or rule out a condition. It may be computed using the equation below [21].

$$\text{Acc}=\frac{TP+TN}{TP+TN+FP+FN}\times 100\% \tag{3}$$

B- (Recall): The sensitivity indicates the percentage of true positives that are accurately detected. The sensitivity may be determined using the equation below [21].

$$\text{Recall} = \frac{\boldsymbol{T_P}}{\boldsymbol{T_P+F_N}} \tag{4}$$

C- (Specificity): As shown in the equation below, this may be calculated by dividing the true negative by the total number of negatives [21].

$$\text{Specificity} =\frac{\boldsymbol{T_N}}{\boldsymbol{T_N+F_P}} \tag{5}$$

d- Precision **(PPV):** This is the likelihood that a patient who gets a positive screening test has the disease. As stated in the equation, the PPV may be calculated [21].

$$\textbf{PPV} =\frac{T_p}{T_p+F_p} \tag{6}$$

e- Negative predictive value (NPV): This reflects the likelihood of discovering a patient who is not at risk for heart disease and is calculated using the equation below [21].

$$\text{NPV} =\frac{T_N}{T_N+F_N} \tag{7}$$

Table 3 shows the confusion matrix generated by the suggested model for three methods.

Table 3-Confusion Matrix Values Obtained Using Various Algorithms

Algorithm	TP	FP	TN	FN
KNN	86	9	260	9
Naïve Bayes	94	7	355	7
SVM	71	30	356	30

Table 4-Classification Algorithms Results

Algorithm	Recall	PPV	Specificity	NPV	Accuracy	F-measure
KNN	0.905	90.52	0.966	96.65	95.1%	1.792
Naïve Bayes	0.93	93.06	0.980	98.06	96.9%	1.841
SVM	0.70	70.92	0.922	92.22	87.4%	1.386

VI. Analysis of Results

The proposed classifiers SVM, KNN, and Naïve Bayes are compared based on the following parameters: testing time, specificity, precision, recall, NPV, and accuracy of the result obtained. This comparison aims to find the best performance classifier among the proposed classifiers to diagnose heart disease The performance comparison results of the classifiers when 65 % of the UCI data is used for training and 35% of the data for testing are shown in Figure 2. The specificity of Naïve Bayes is comparatively high compared with other classifiers, which is 98.0%. It can be noted that the lowest accuracy has been recorded by the

SVM classifier (87.4%), followed by KNN (95.1%) while Naïve Bayes has the best performance with 96.9%. Also, the Naïve Bayes is faster than the other classifiers at getting the diagnosis results. Moreover, in Naïve Bayes, the NPV is higher (98.06) than in SVM (92.22) and (96.65) KNN classifiers. As for Recall and Precision, KNN records (90.5) in recall and (90.5) in precision. Naïve Bayes produced 93.0 and 93.06 for recall and precision, respectively. Finally, the SVM classifier performance was the worst with (70.0) and (70.92) for recall and precision, respectively. To summarize, the Naïve Bayes was the most robust classifier among the proposed classifiers. Figure 3 shows the confusion matrix for the Naive Bayes classifier.

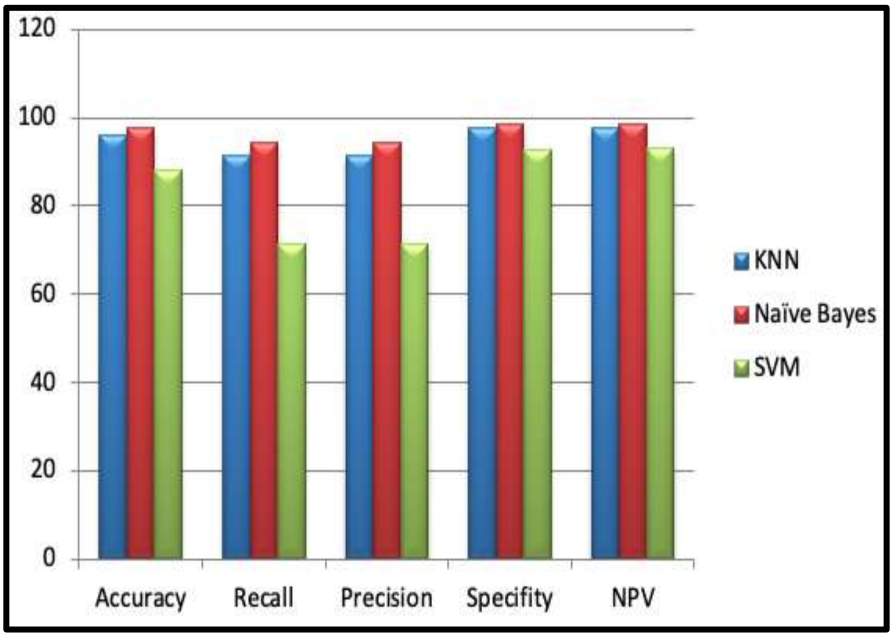

Figure 2- the performance comparison of three algorithms

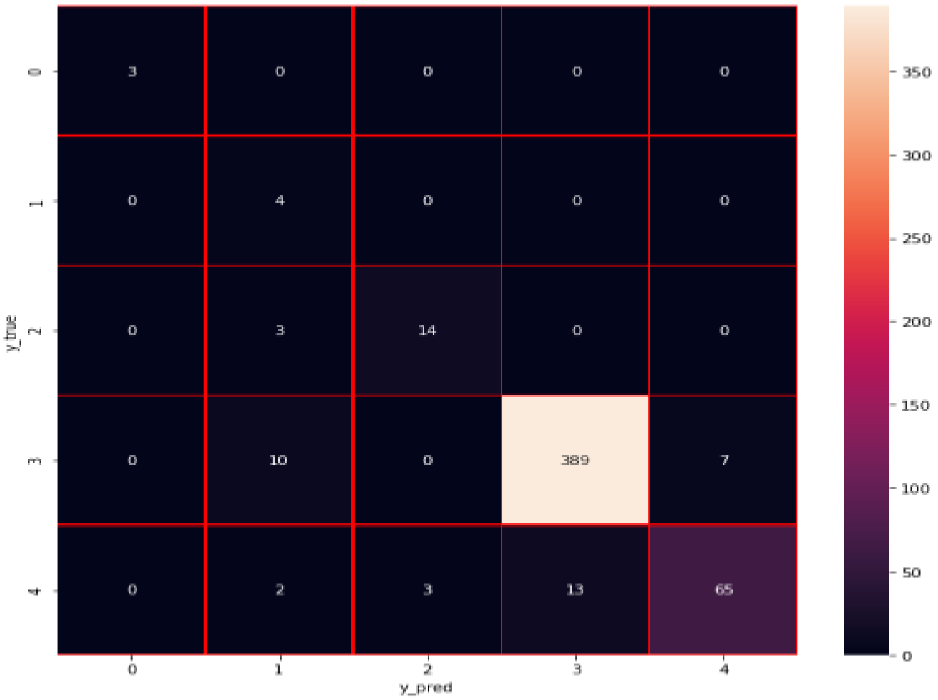

Figure 3-Confusion Matrix of Naïve Bayes

VII. CONCLUSION

With the increasing number of deaths due to heart diseases, it has become necessary to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient machine learning algorithm for the detection of heart diseases. This study compares the accuracy scores of KNN, Naive Bayes, and Support Vector Machine algorithms for the classification of heart diseases using the UCI machine learning repository dataset after preprocessing it. The result of this study indicates that the Naive Bayes algorithm is the most efficient algorithm with an accuracy score of 96.9% for the prediction of heart disease. The main limitation encountered in this work is the inability to diagnose other types of heart diseases such as (Heart Valve Disease, Pericarditis, Wolff-Parkinson-White syndrome, and Congenital heart disease), as the diagnosis of these diseases depends on other features not available in the UCI database. The work can be enhanced in the future by developing a web application based on the Naïve Bayes algorithm as well as by using a real database that contains more features that help classify other heart diseases.

REFERENCES

[1] A. L. Bui, T. B. Horwich, and G. C. Fonarow, “Epidemiology and risk profile of heart failure,” *Nat. Rev. Cardiol.*, vol. 8, no. 1, pp. 30–41, 2011.

[2] P. A. Heidenreich et al., “Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association,” *Circulation*, vol. 123, no. 8, pp. 933–944, 2011.

[3] M. Durairaj and N. Ramasamy, “A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate,” *Int. J. Control Theory Appl,* vol. 9, no. 27, pp. 255–260, 2016.

[4] J. Mourao-Miranda, A. L. W. Bokde, C. Born, H. Hampel, and M. Stetter, “Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data,” *Neuroimage*, vol. 28, no. 4, pp. 980–995, 2005.

[5] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, “Innovative artificial neural networks-based decision support system for heart diseases diagnosis,” 2013.

[6] Q. K. Al-Shayea, “Artificial neural networks in medical diagnosis,” *Int. J. Comput. Sci*. Issues, vol. 8, no. 2, pp. 150–154, 2011.

[7] J. López-Sendón, “The heart failure epidemic,” *Medicographia*, vol. 33, no. 4, pp. 363–369, 2011.

[8] K. Vanisree and J. Singaraju, “Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks,” Int. J. Comput. Appl., vol. 19, no. 6, pp. 6–12, 2011.

[9] A. Methaila, P. Kansal, H. Arya, and P. Kumar, “Early heart disease prediction using data mining techniques,” Comput. Sci. Inf. Technol. J., vol. 24, pp. 53–59, 2014.

[10] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, “An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction,” Expert Syst. Appl., vol. 68, pp. 163–172, 2017.

[11] R. Detrano et al., “International application of a new probability algorithm for the diagnosis of coronary artery disease,” Am. J. Cardiol., vol. 64, no. 5, pp. 304–310, 1989.

[12] U. Raghavendra et al., “Automated technique for coronary artery disease characterization and classification using DD-DTDWT in ultrasound images,” *Biomed. Signal Process. Control,* vol. 40, no. October, pp. 324–334, 2018, doi: 10.1016/j.bspc.2017.09.030.

[13] F. S. Alotaibi, “Implementation of machine learning model to predict heart failure disease,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 6, pp. 261–268, 2019.

[14] Latha, C.B.C.; Jeeva, S.C. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques.Inform. Med. Unlocked 2019, 16, 100203. [CrossRef]

[15] Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access 2019, 7, 81542–81554. [CrossRef]

[16] Louridi, N.; Amar, M.; El Ouahidi, B., "Identification of Cardiovascular Diseases Using Machine

Learning," In Proceedings of the 7th Mediterranean Congress of Telecommunications 2019, CMT 2019, Fez, Morocco, 24–25 October 2019; pp. 1–6. [CrossRef]

[17] Gupta, A.; Kumar, R.; Arora, H.S.; Raman, B., "MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis," *IEEE Access* 2019, 8, 14659–14674. [CrossRef]

[18] Perumal, R., "Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques," *Int. J. Adv.Sci. Technol*. 2020, 29, 4225–4234.

[19] Kumar, N.K.; Sindhu, G.; Prashanthi, D.; Sulthana, A., "Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers," In Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems(ICACCS), Coimbatore, India, 6–7 March 2020; pp. 15–21. [CrossRef]

[20] Gazelo ˘glu, C., "Prediction of heart disease by classifying with feature selection and machine learning methods," *Prog. Nutr*. 2020,22, 660–670. [CrossRef]

[21] V. Sharma, S. Yadav, and M. Gupta, “Heart Disease Prediction using Machine Learning Techniques,” in 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 177–181.

[22] Pavithra, V.; Jayalakshmi, V., "Hybrid feature selection technique for prediction of cardiovascular diseases," Mater. Today Proc. 2021,22, 660–670. [CrossRef]

[23] Kavitha, M.; Gnaneswar, G.; Dinesh, R.; Sai, Y.R.; Suraj, R.S., "Heart Disease Prediction using Hybrid machine Learning Model," In Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021, Coimbatore, India, 20–22 January 2021; pp. 1329–1333. [CrossRef]

[24]B. S. Kumar, “A Survey on Data Mining Techniques for Prediction of Heart Diseases,” IOSR J. Eng. www. iosrjen. org ISSN, vol. 8, pp. 22–27, 2018.

[25] S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K. Bashir, “Improving heart disease prediction using feature selection approaches,” in 2019 16th international bhurban conference on applied sciences and technology (IBCAST), 2019, pp. 619–623.

[26] F. Razaque et al., “Using naïve bayes algorithm to students’ bachelor academic performances analysis,” in 2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS), 2017, pp. 1–5.

[27] C. B. Rjeily, G. Badr, A. H. El Hassani, and E. Andres, “Medical data mining for heart diseases and the future of sequential mining in medical field,” in Machine Learning Paradigms, Springer, 2019, pp. 71–99.

[28] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, “Sentiment analysis of review datasets using naive bayes and k-nn classifier,” arXiv Prepr. arXiv1610.09982, 2016.

[29] O. Qasim and K. Al-Saedi, “Malware Detection using Data Mining Naïve Bayesian Classification Technique with Worm Dataset,” Int. J. Adv. Res. Comput. Commun. Eng, vol. 6, no. 11, pp. 211–213, 2017.

[30] I. Babaoğlu, M. S. Kıran, E. Ülker, and M. Gündüz, “Diagnosis of coronary artery disease using artificial bee colony and k-nearest neighbor algorithms,” Int. J. Comput. Commun. Eng., vol. 2, no. 1, pp. 56–59, 2013.

[31] S. K. Kotha, J. Pinjala, K. Kasoju, and M. Pothineni, “Gesture Recognition System,” Int. J. Res. Eng. Technol., 2015.

[32] P. Tabesh, G. Lim, S. Khator, and C. Dacso, “A support vector machine approach for predicting heart conditions,” in IIE Annual Conference. Proceedings, 2010, p. 1.

[33] P. Tabesh, G. Lim, S. Khator, and C. Dacso, "A support vector machine approach for predicting heart conditions, in Proceedings of the 2010 Industrial Engineering Research Conference, 2010, p. 5.

[34] UCI, "Heart Disease Data Set." [Online]. Available: https://www.kaggle.com/ronitf/heart-disease-uci. [Accessed: 20-Apr2019].

[35] E. Acuna, “preprocessing in Data Mining,” Int. Encycl. Stat. Sci., no. September, 2011, DOI: 10.1007/978-3-642-04898-2.

[36] Y. Jung and J. Hu, “A K-fold averaging cross-validation procedure,” J. Nonparametr. Stat., vol. 27, no. 2, pp. 167–179, 2015, DOI: 10.1080/10485252.2015.1010532.