International Journal of Engineering & Technology, 10 (1) (2021) 46-56

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Heart Disease Prediction Model Using Naïve Bayes Algorithm and Machine Learning Techniques

Maria Yousef [1]*, Prof. Khaled Batiha[2]

[1] Department of Computer Science, Al al-Bayt University, Jordan
[2]Professor, Department of Computer Science, Al al-Bayt University, Jordan
**Corresponding author E-mail: maria.yousef@yahoo.com*

Abstract

These days, heart disease comes to be one of the major health problems which have affected the lives of people in the whole world. Moreover, death due to heart disease is increasing day by day. So the heart disease prediction systems play an important role in the prevention of heart problems. Where these prediction systems assist doctors in making the right decision to diagnose heart disease easily. The existing prediction systems suffering from the high dimensionality problem of selected features that increase the prediction time and decrease the performance accuracy of the prediction due to many redundant or irrelevant features. Therefore, this paper aims to provide a solution of the dimensionality problem by proposing a new mixed model for heart disease prediction based on (Naïve Bayes method, and machine learning classifiers).

In this study, we proposed a new heart disease prediction model (NB-SKDR) based on the Naïve Bayes algorithm (NB) and several machine learning techniques including Support Vector Machine, K-Nearest Neighbors, Decision Tree, and Random Forest. This prediction model consists of three main phases which include: preprocessing, feature selection, and classification. The main objective of this proposed model is to improve the performance of the prediction system and finding the best subset of features. This proposed approach uses the Naïve Bayes technique based on the Bayes theorem to select the best subset of features for the next classification phase, also to handle the high dimensionality problem by avoiding unnecessary features and select only the important ones in an attempt to improve the efficiency and accuracy of classifiers. This method is able to reduce the number of features from 13 to 6 which are (age, gender, blood pressure, fasting blood sugar, cholesterol, exercise induce engine) by determining the dependency between a set of attributes. The dependent attributes are the attributes in which an attribute depends on the other attribute in deciding the value of the class attribute. The dependency between attributes is measured by the conditional probability, which can be easily computed by Bayes theorem. Moreover, in the classification phase, the proposed system uses different classification algorithms such as (DT Decision Tree, RF Random Forest, SVM Support Vector machine, KNN Nearest Neighbors) as a classifiers for predicting whether a patient has heart disease or not. The model is trained and evaluated using the Cleveland Heart Disease database, which contains 13 features and 303 samples.

Different algorithms use different rules for producing different representations of knowledge. So, the selection of algorithms to build our model is based on their performance. In this work, we applied and compared several classification algorithms which are (DT, SVM, RF, and KNN) to identify the best-suited algorithm to achieve high accuracy in the prediction of heart disease. After combining the Naive Bayes method with each one of these previous classifiers the performance of these combines algorithms is evaluated by different performance metrics such as (Specificity, Sensitivity, and Accuracy). Where the experimental results show that out of these four classification models, the combination between the Naive Bayes feature selection approach and the SVM RBF classifier can predict heart disease with the highest accuracy of 98%. Finally, the proposed approach is compared with another two systems which developed based on two different approaches in the feature selection step. The first system, based on the Genetic Algorithm (GA) technique, and the second uses the Principal Component Analysis (PCA) technique. Consequently, the comparison proved that the Naive Bayes selection approach of the proposed system is better than the GA and PCA approach in terms of prediction accuracy.

Keywords: *Heart Disease; Naïve Bayes; Bayes Theorem; Feature Selection; Prediction; Accuracy.*

1. Introduction

The main objective of hospitals is to provide high levels of health care and good treatment services within the best of their potentials and qualities. Health care includes a specific set of basic services provided by institutions in both the public and private sectors. They present treatment for health problems as well as disease prevention plans and also improvement of health behaviors according to the patient's situation. This health care system includes hospitals, pharmacies, and clinics, places with human medical cadre's workers such as doctors, nurses along with workers in the field of medical research, people having the aim of providing citizens with health care [1].

Health care is provided to all patients in hospitals to achieve the general benefit. The benefit of high quality because it is related to human life. It should be direct contact between the beneficiary of health care services and the hospital. Health care service can only be provided

in the presence of the patient himself for examination, analytical procedure, and diagnosis of treatment but there is a delay in the provision of health services [2].
The hospital database has a huge amount of data such as patient information, and health information. Where, patient's information includes (patient number, patient name, patient age, address, date, gender), and health information includes (date, doctor's number, patient number, patient type, doctor's report, examination number, prescription number, type of medication), in addition to data on the drug itself (name of medicine and quantity of medicine) [3]. Moreover, the medical historical record of the patient helps the doctor to follow up on the patient's situation by going through his files, which contain general medical tests, results, and treatment prescriptions. Those patient's files must be available for the doctor in the hospital database [4].
Usually, the disease is diagnosed based on the experience of the doctor, rather than using useful information gained from the hospital database. Sometimes, doctors may fail to assume correct decisions when diagnosing the disease. So the treatment may include bias and errors. This may also result in a negative effect on the patient's health. Presently, doctors are using many scientific technologies such as automatic prediction systems for both identification and diagnosing several types of diseases.
Successful treatment is always associated with a right and accurate diagnosis. Cleveland Heart Disease database consists of 13 features that described the medical information of the patients, as some of these features are irrelevant and not important to the prediction process. Most of the existing heart disease prediction systems are depending on all features in the database and uses it to predict if a patient is suffering from heart disease or not, without giving importance to reducing the number of features and choosing the best one for the prediction. This leads to the appearance of a high dimensionality problem that may reduce the prediction results and reduce the performance accuracy of prediction.
The principal contribution of this paper will be to propose a mixed algorithm (NB-SKDR) as a new classifier for heart disease prediction in the aims to address the high dimensionality problem of features and improve the prediction accuracy. We implementing our proposed system by combining the Naive Bayes (NB) feature selection algorithm with different classification algorithms (DT, SVM, RF, and KNN) and compare their effect on accuracy and classification performance.
The rest of the paper is organized as follows. Section 2 reviews some of the related work that used different classification techniques to classify heart disease. Section 3 contains the steps for implementation of the proposed heart disease prediction model. Section 4presents the performance results of the proposed system and compares the results with some other systems in the same
field. Section 5 concludes this study.

2. Related work

In this section, we will discuss an overview of the previous studies that proposed various systems to predict the problem of heart disease by using various classification techniques such as (Decision Tree, Random Forest, Nearest Neighbor, and Support Vector Machine). Moreover, several techniques of feature selection have been presented in this section.
In [5], the authors have presented a new model for predicting heart disease using data mining techniques to predict the likelihood of a patient getting a heart disease. In this research, the supervised machine learning algorithm (Naive Bayes, KNN, and DT) was applied on the Cleveland Heart Disease dataset which represents the historical medical files and contains 303 records with 13 attributes. The performance of the classifiers is evaluated and their results achieved 52.3% of accuracy when using the Naive Bayes classifier, while DT and KNN obtained 52%, 45.6% of accuracy respectively.
Authors in [6] introduced a paper in which they implemented DT, SVM, RF, Naive Bayes, and Logistic Regression for the prediction of heart disease. The main purpose of this study was to help doctors with the prediction by comparing the prediction accuracy of different machine learning algorithms. The author depends on the heart disease database which includes 303 records and 13 attributes. After experimentation, it was concluded that SVM and RF obtained 84%, 91% of accuracy respectively, both algorithms generated better results compared to another algorithm.
In the paper presented by [7], the authors proposed a new hybrid methodology for the classification and prediction of heart disease based on GA (Genetic Algorithm) and SVM RBF supervised learning technique. The GA is used as a feature reduction algorithm to identify the best subset of features by select “N best features for classification out of total M features” to improve the SVM RBF classification with high accuracy. This hybrid approach has been applying on the Cleveland Heart Disease database to reduce the features from 13 to 7. The classification result achieved by this study is 88.1% of accuracy for the diagnosis of a disease.
In this paper [8], the authors developed a new heart disease prediction model by combining two different algorithms which are Principal Component Analysis (PCA) and Support Vector Machine (SVM) with the aim to achieve improved accuracy for the prediction. The researchers developed this model in two stages. In the first stage, the PCA technique was used as a reduction algorithm to select relevant features from the Cleveland Heart Disease dataset, where it was able to reduce the features from 13 to 6 features that are essentially more related to the diagnosis of heart disease. In the second stage, the authors used the SVM RBF algorithm as a classifier for the prediction process. The proposed model has shown 54.3% accuracy in predicting.
In another study conducted by [9], the authors developed a hybrid heart disease prediction system that helps the doctors to diagnose the patient based on patient conditions. The Hybrid classification approach includes two algorithms (KNN and DT), the KNN algorithm is applied for prediction analysis, and a decision tree is applied for classification. The result of classification shows 86% of accuracy by using DT and KNN with n number of neighbors. The main limitation of this study that the accuracy of hybrid KNN and DT is not enough.
In this study [10], the authors attempt to specify the most efficient data mining technique used for medical prediction purposes by reviewed 168 articles associated with the implementation of data mining for medical prediction systems between 1997 and 2018. The authors selected 85 of these studies identified by using databases like IEEE Explore, Google Scholar, Science Direct, and Springer Link. After analysis of all selected studies, the authors demonstrate that the hybrid method provides better prediction accuracy results than using one algorithm as alone such as the study of [11]. So using the hybrid approach is recommended in future studies.
Another attempt to provide a comparative study of several machine learning algorithms is done by [12]. The authors apply (KNN, DT, and SVM) algorithms on the Cleveland Heart Disease database to compare the performance of the prediction. The results of the experiments indicate that the KNN gives an accuracy of 83.1% when the value of K equals 9, while DT and SVM achieved 82.1%, 84.1% of accuracy respectively.

3. Methodology

This section provides the main phases of the proposed prediction model for classifying heart disease. Important concepts such as the data set, classification algorithms, and data mining techniques are described following.

3.1 Database description

The dataset used in this study to build our proposed system is the Cleveland Heart Disease dataset. This dataset is obtained from the University of California, Irvine (UCI) machine learning repository and available online at: (http://archive.ics.uci.edu/ml/datasets/Heart+Disease).This dataset contains 303 samples (rows) for patients who are suffering from the disease or not, and a total number of 76 features (columns) to describe the medical information of patients such as analyzes and medical examinations. But all published experiments among them [13] and [9] refer to using only 14 of those features which are closely linked to heart disease, it was observed that heart disease prediction was only dependent on 14 features that affect the incidence of heart disease. Moreover, The Cleveland Heart Disease dataset free from noisy data and contains very little null data. So, it is one of the most commonly used databases for heart disease prediction systems. The database attributes used in this study are described in Table 1. Also, Fig.1 displays part of the heart disease database used.

Table 1: Description of the Cleveland Heart Disease dataset

Attribute	Description	Value
Age	displays the age of the individual in year	Child (0-12 years) Adolescence (13-18 years) Adult (19-59 years) Senior Adult (60 years and above)
Sex	displays the gender of the individual	Male=1 Female=0
CP	displays the type of chest-pain experienced by the individual	Typical=1 Atypical=2 Non-anginal pain=3 Asymptomatic=4
Trestbps	displays the resting blood pressure value of an individual in mmHg(unit)	(less than 120)/(less than 80) = 0: Ideal (120-129)/(80-84) = 1: Normal (130-139)/(85-89)= 2 :Up-Normal (Higher than 140)/(Higher than 90)= 3 Higher(Stage1,Stage2,Stage3)
Chol	Serum Cholesterol in mg/dl	0 = Optimal < (150) 1= Desirable (150 – 199) 2 = High (200 – 399) 3 = Very high ≥ (400)
Fbs	Fasting Blood Sugar > 120 mg/dl)	1 = true / 0 = false
restecg	Resting Electrocardiograph Results	0 = normal 1 = having ST-T wave abnormality 2 = left ventricular hyperthrophy
thalach	Displays the max heart rate achieved by an individual.	0= Low <111 1= Medium 111 – 141 2= High >141
exang	Exercise induced angina	1 = Yes / 0 = No
oldpeak	ST depression induced by exercise relative to rest, displays the value which is an integer or float.	Domain:0- 6.2
Slope	The Slope of The Peak Exercise ST segment	1= Up Sloping 2= Flat 3=Down Sloping
Ca	Number of major vessels (0-3) colored by fluoroscopy	0 = 8 Vessels 1= 3 Vessels 2 = 2 Vessels 3 = 1 Vessels
thal	Displays the Thalassemia Test	3 = Normal 6 = Fixed defect 7 = Reversible defect
target	Displays whether the individual is suffering from disease or not	0= Uninfected / 1= infected

Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
58	0	3	150	283	1	0	162	0	1	2	0	2	1
50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
58	0	2	120	340	0	1	172	0	0	2	0	2	1
66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
42	1	0	140	226	0	1	178	0	0	2	0	2	1

Fig. 1: Part of the Cleveland Heart Disease database

3.2 Technical background

In this section, we will introduce a description of different classification methods used for diagnoses of heart disease as the following:

Decision Tree Algorithm (DT): A Decision Tree is a graphical method that is frequently used to visually and explicitly represent the decision, it can be used to solve both regressions and classification problems. A decision tree includes decision nodes, branches, and leaf nodes. Each decision node represents a feature (attribute), also each branch presents sequence rules that can be used to classify the data, as well as each leaf node matches to a class label that represents the decision result. To build a Decision tree algorithm we need to perform heuristic measures that can lead us to the best possible split of the tree. Entropy and Information Gain are two of those heuristic measures where entropy is used for controls how a decision tree decides to split the data, it can be calculated as shown in equation (1) according to [14]. While the Information Gain (IG), measures how much information a feature gives us about the specific class we will use it to decide the ordering of attributes in the nodes of a decision tree, it can be defined using equation (2) according to [15].

$$\boldsymbol{E} = \sum_{i}^{n} -\boldsymbol{p_i log_2 p_i} \tag{1}$$

Where p_i the probability/percentage of samples, that belong to class i for a particular node.

$$\boldsymbol{G} = \boldsymbol{E(parent)} - \boldsymbol{Average\ E(children)} \tag{2}$$

Where $Average\ E()$ mean the sum of weighted entropies of each branch (children) from the parent node.

Support Vector Machine (SVM): SVM is a type of supervised machine learning algorithm and a widespread method for many machine learning tasks, also it is a useful technique for classification both linear and non-linear data, and it is proposed by [16]. When the data is linearly separable, the task of SVM is to find the optimal hyperplane line "decision boundary" that separates the data in two classes (to the left, and the right) of the hyperplane line. On the other hand, it is not possible to find a hyperplane for the non-linear separation problem. So, the kernel function is used to transform the data from low dimensional into a higher dimensional in aims to make the data more suitable for classification purposes as explained by [17]. SVM algorithm uses different types of Kernel Functions to map the original dataset (linear or nonlinear) into a higher dimensional space in an attempt to make it as a linear dataset. Linear, Nonlinear, Polynomial, Radial Basis Function (RBF), and Sigmoid are different types of kernel functions. The most used type is RBF, and it can be defined by equation (3) according to [18].

$$\boldsymbol{K(x, x')} = \boldsymbol{exp}\ (-\boldsymbol{\gamma}||\boldsymbol{x} - \boldsymbol{x'}||^2) \tag{3}$$

Where x and x' are two samples represented as feature vectors in some input space. While γ is a parameter that sets the "spread" of the kernel. $\gamma = \frac{1}{2\sigma^2}$

Nearest Neighbor Algorithm (KNN): This is a powerful classification algorithm used for performing both classification and regression predictive problems. KNN is a none- parametric and lazy learner algorithm because it does not have a specialized training phase it always memories all training datasets and uses it when predicts a new data point [19] Besides, the KNN algorithm reaches its decision based on similarity measures to calculate the distance between a new sample to predict and existing training samples, the training sample that has a shorter distance to the new one is more likely to have the same class label. Euclidean Distance is one of the most used distance metrics to calculate the distance between two points [20], and it is given by equation (4).

$$\boldsymbol{d(p, q)} = \sqrt{\boldsymbol{(q_i - p_i)^2}} \tag{4}$$

Where q_i is the coordinate of the first point, and p_i is the coordinate of the second point.

Random Forest Algorithm (RF): it is a composite classifier consisting of many decision trees and used for both classification and regression tasks. Furthermore, the random forest algorithm builds several decision trees on different random samples from a specific dataset [21]. In this algorithm, each one of the decision trees can be built separately on a subset of the data using the Entropy and Information Gain heuristic to get a final result (decision). Typically, the random forest algorithm is based on a voting technique that selects the most voted prediction result as the final decision [22].

Naïve Bayes Algorithm (NB): This algorithm is a powerful probabilistic representation that can be used to create models with the predictive capability to display the probability of each input attribute for the predictable state. This algorithm relies on the Bayes Theorem which describes the probability of a feature based on prior knowledge of conditions that might be related to that feature [23]. Naive Bayes algorithm is considered to be "naïve" because the independence between every pair of features to given the value of the class variable is the major assumption to make a prediction. Although the Naive Bayes (NB) algorithm is simple, it is very effective in many real world datasets because it can give better predictive accuracy.

3.3 The proposed (NB-SKDR) heart disease prediction system

In this research, a mixed heart disease decision support system (NB-SKDR) is introduced. The proposed system is based on the Naive Bayes technique (NB) to select the best subset of features and four machine learning algorithms Namely (DT, RF, SVM, and KNN) as a classifier. The proposed (NDRV-NB) system contains three main phases including preprocessing, feature selection, and classification phase. The main objective of this system is to identify the best subset of features for classification and optimize the accuracy of the heart disease prediction system. The proposed system framework is presented in Fig.2. Also, the proposed system phases are described in the following subsection.

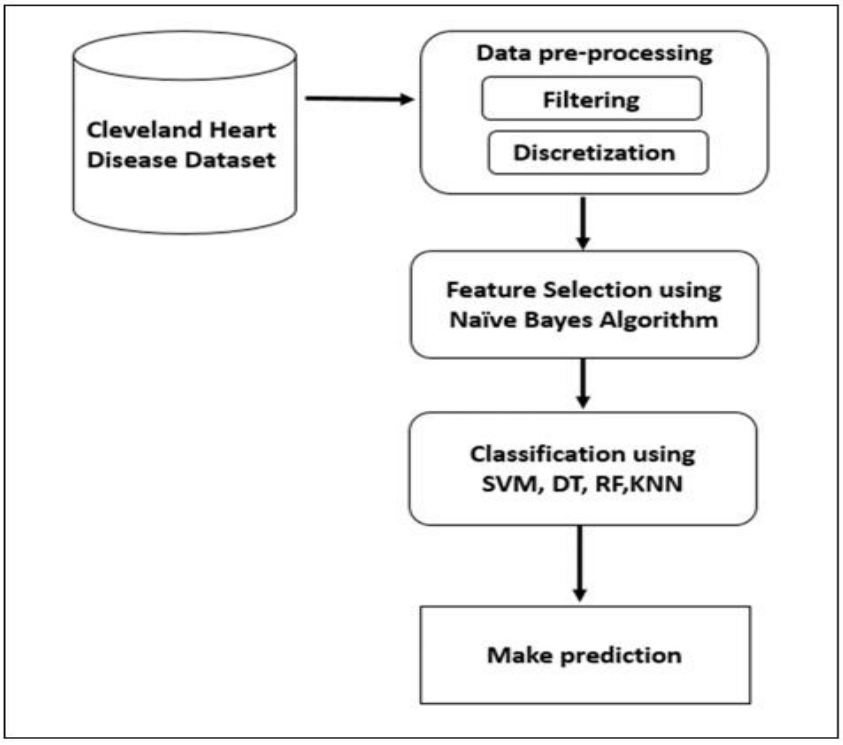

Fig. 2: The proposed (NB-SKDR) heart disease prediction model framework.

3.3.1 The preprocessing phase

Data preprocessing is one of the data mining matters and considered to be one of the top main phases of the prediction process. The primary objective of data preprocessing is to prepare and transform the raw data that may have incomplete records, noise value, outlier, and inconsistent data into a useful and suitable format. It includes several strategies like Data Cleaning, Data Transformation, and Data Reduction strategy. Moreover, Any Decision Support System which handles a very large volume of data requires effective Data Preprocessing to enhance the input dataset reliability and availability before running and analysis as demonstrated by [24].
In this proposed system, the preprocessing phase achieves through several steps such as filtering and discretization operation. Where both operations have been applied on the Cleveland heart disease dataset. The null data are firstly solved by pass the dataset through a filtering technique that replaces all null values with a value inferred from the column values. Then, the dataset was discretized into the proper format for the feature selection and classification process. The data preprocessing operations are illustrated in the following subsections.

- **Filtering Technique**: Any null values that existed in the dataset negatively effect on the feature selection and later on the classification process. Also, null data causes two primary problems. First, leads to loss of precision due to fewer data. Where the second problem is a computational difficulty due to the invalid value in the dataset. The training dataset used for this experiment contained 303 instances, out of which 6 records contain some null value in the attribute Number of major vessel colors by fluoroscopy (Ca), and Thalassemia test (thal), as shown in Fig.3. So, there is needed to handle null data by using the filtering technique. This technique is applied on a Cleveland dataset to find out and solve null data problems through several ways described by [25].

```
Out[2]:
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          4
thal        2
target      0
dtype: int64
```

Fig.3: Described how many null values in our heart disease dataset

From the previous figure, it was found that there are 6 null values in the two features (ca and thal). In our proposed system, the filtering method evaluates the mean of every attribute in which the data was null, and this mean value is replaced at the null value place. Where the mean is calculated according to the following equation (5). By using the Filtering method, null data will be removed from the Cleveland Heart Disease dataset.

$$\mu = \frac{\sum_i^n \boldsymbol{x_i}}{\boldsymbol{n}} \quad (5)$$

Where x_i denotes the value for each instance and n denotes the total frequency.

- **Discretization:** Data Transformation is an approach to convert the original value into the appropriate format in an attempt to make the dataset well-structured and suitable for the next classification step. In this phase, we applied the data transformation technique it's called "Discretization" as a scale to convert data from a large domain of numeric values to a subset of categorical values before it becomes used for the selection and classification stage. Because the Cleveland heart disease database contains medical data in the various numeric ranges such as Trestbps and Chol, the discretization step is particularly used to manage and organize the values of features in the dataset [26]. This routine is applied to discretize numerical features by partitioning the value of the feature into groups or clusters, after that each raw value of the numeric feature is replaced by an interval label. Additionally, Fig.4 shows the Cleveland Heart Disease database after discretization preprocessing.

Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	target
3	1	3	3	2	1	0	2	0	0	1	1		
2	1	2	2	2	0	1	2	0	3.5	0	0	2	1
2	0	1	2	2	0	0	2	0	1.4	2	0	2	1
2	1	1	1	2	0	1	2	0	0.8	2	0	2	1
2	0	0	1	2	0	1	2	1	0.6	2	0	2	1
2	1	0	3	1	0	1	2	0	0.4	1	0	1	1
2	0	1	3	2	0	0	2	0	1.3	1	0	2	1
2	1	1	1	2	0	1	2	0	0	2	0	3	1
2	1	2	3	1	1	1	2	0	0.5	2	0	3	1
2	1	2	3	1	0	1	2	0	1.6	2	0	2	1
2	1	0	3	2	0	1	2	0	1.2	2	0	2	1
2	0	2	2	2	0	1	2	0	0.2	2	0	2	1
2	1	1	2	2	0	1	1	0	0.6	2	0	2	1
3	1	3	0	2	0	0	2	1	1.8	1	0	2	1
2	0	3	3	2	1	0	2	0	1	2	0	2	1
2	0	2	1	2	0	1	2	0	1.6	1	0	2	1
2	0	2	1	2	0	1	2	0	0	2	0	2	1
3	0	3	3	2	0	1	2	0	2.6	0	0	2	1
2	1	0	3	2	0	1	1	0	1.5	2	0	2	1
3	0	3	3	2	0	1	2	0	1.8	2	2	2	1
2	1	0	2	2	0	1	2	0	0.5	1	0	3	1
2	1	2	2	2	0	1	2	1	0.4	2	0	2	1
2	1	0	3	2	0	1	2	0	0	2	0	2	1

Fig.4: Cleveland heart disease database after discretization preprocessing.

3.3.2 Features Selection Phase

The feature selection step is considered an important phase in the proposed prediction system to solve the high dimensionality problem of the feature set. This technique eliminates irrelevant and redundant features by search in the dataset and selects the most necessary and sufficient subset of features to achieve similar or even better classification performance than using all features because not all features in the dataset are beneficial and have a strong association with prediction process as shown in [27]. The purpose of the feature selection procedure is to provide a faster and more cost effective predictor due to minims the number of features, it can also improve the classification accuracy due to select the best subset of features [28].

This research employed the Naive Bayes algorithm as one method of feature selection based on the Bayes Theorem. The main idea of this algorithm is to reduce the dimensionality of the attributes by identifying the most important attributes which are sufficient for the classification task and avoid unnecessary Attributes. The Naïve Bayes algorithm performs by determining the dependency between a set of

attributes. The dependent attributes are the attributes in which an attribute depends on the other attribute in deciding the value of the class attribute [29]. The dependency between attributes is measured by the conditional probability, which can be easily computed by Bayes Theorem.

Accordingly, the Naive Bayes method was applied on the Cleveland heart disease dataset which contains 13 features to extract a new subset includes 6 dependent features which are (age, gender, blood pressure, fasting blood sugar, cholesterol, exercise induce engine). Dependency between features is evaluated by first grouping them, then using the Bayes theorem to calculate the probabilities of their values in determining the value of the class feature, and computed the difference between the probabilities, this procedure is repeated for all acceptable combinations of features. Finally, the dependencies between particular features in a total features set are found. For more description, this method is accomplished as the following steps:

Step 1: Let the initial set of attributes be $A = \{a_1, a_2, a_3, \dots, a_n\}$ where feature a_1 can take values $=\{a_{11}, a_{12}, a_{13}, \dots, a_{1n}\}$, a_2 can take values$=\{a_{21}, a_{22}, a_{23}, \dots, a_{2n}\}$, a_n can take values$=\{a_{n1}, a_{n2}, a_{n3}, \dots, a_{nn}\}$, and class attribute $C\ can\ take\ values = \{c_1, c_2\}$.

Step2: Combination between attributes, by group the attributes in set A into an attributes sets of pairs A$=\{A_0, A_1, A_2, \dots, A_n\}$, where $A_0 = \{a_1, a_2\},\ A_1 = \{a_1, a_3\},\ A_n\{a_1, a_n\}$.

Step3: Find the dependency of all pairs of features in determining the value of the class feature. The dependency of two features is calculated by the conditional probabilities of the class feature given the values of the features, which can be easily measured by Bayes theorem as the following equation (6).

$$\boldsymbol{P(c|feature_1, feature_2)} = \frac{\boldsymbol{P(c)\ P(feature_1|c)\ (feature_2|c)}}{\boldsymbol{P(feature_1) + P(feature_2)}} \quad (6)$$

Where:
$P(c|feature_1, feature_2)$, is the conditional probability of class c, given $feature1$ and$feature2$. To determine if class c derived from or depends upon the specified value of $feature1$ and $feature\ 2$.
$P(feature_1|C)$, is the conditional probability of $feature1$ given class c.
$P(feature_1),\ P(feature_2)$, is the prior probability or marginal probability of $feature1$ and $feature\ 2$. It is "prior" in the sense that it does not take into account any information about class c.

The Naïve Bayes algorithm

1-[Input]: Original dataset A with all attributes

2-[Group]: Divide A in subset each pairs of attribute together // $A_0 = \{a_1, a_2\},\ A_1 = \{a_1, a_3\}, A_2 = \{a_1, a_4\},\ A_n = \{a_1, a_n\}, etc.$

3-[Initialize B]: new subset to store independent features // B=Ø

4-[Estimate Probability]: Naïve Bayes based on By Bayes Theorem rule

```
When i=0;
Get ( A_i ) ; // subset contain pairs of attribute
For (j=1, j<=n, j++) // n equal the distinct number of value in class attrubite
{
P ( A_i[0], A_i[1]|CL_j); // CL present value of class attribute,
A_i[0] first feature, A_i[1] second feature in the A_i subset
}
IF dependency exists THEN
Store the attribute A_i[0], A[1] into another subset B;
Else increase i;
```

5-[Output]: subset of Attributes B // contains dependent Attributes

Step4: Now, after the relationship between features is discovered, the dependent features are stored in a new subset B. Where, the B subset represents the final subset that contains the most important features for the next classification phase.

3.3.3. Classification phase using (RF, SVM, DT, KNN)

A supervised machine learning algorithms has crucial importance in medical prediction and widely apply in various disease classification problems to predict the class of objects whose class label is unknown [30].
In this phase, the proposed (NDRV-NB) model is developed based on four classification techniques which are (Support Vector Machine, Decision tree, Neighrest Neighbor, and Random Forest) to solve the heart disease classification problem. Different algorithms use different rules for producing different representations of knowledge. So, the selection of algorithms to build our system is based on their performance. In this work, we attempt to find the most suitable machine learning techniques for predicting heart patients in the following two steps.

- **First**, the dataset is divided into two subsets (training and testing). We used 65% of data for training and 35% of data for testing. In training, the data is used to implement the proposed model, while testing is used to evaluate the performance of the model and ensure the results.
- **Second**, each one of these algorithms (SVM, DT, RF, and KNN) is used separately to training the model on how to make the prediction and take a decision regarding a patient. The algorithms are based on the new database of 6 features instead of 13 in training and evaluating the model. Finally, we use the algorithm that achieved a higher level of accuracy for the classification process.

4. Experimental Results and Discussion

The system is tested on the Cleveland Heart Disease dataset to assess the model validity and ensure of prediction rate by using several criteria such as:

1) **Accuracy:** is the ratio of the right prediction achieved by using the proposed method. It is calculated by equation (7).

$$\frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{7}$$

2) **Sensitivity:** (recognition rate or true positive rate), the classification probability of the sick patient, means that the ability of a test to correctly identify those with the disease. It is calculated by the equation (8).

$$\frac{TP}{(TP + FN)} \tag{8}$$

3) **Specificity**: (true negative rate), It is used for measuring the percentage of healthy people who are correctly identified from the dataset. It is calculated by the equation (9).

$$\frac{TN}{(TN + FP)} \tag{9}$$

Where:
True positive (TP): The number of positively labeled classes, which have been classified as positive "correctly predicts positive class".

True Negative (TN): The number of negatively labeled classes, which have been classified as negative "correctly predicts the Negative class".

False positive (FP): The number of negatively labeled classes, which falsely have been classified as Positive "incorrectly predicts the Positive class.

False Negative (FN): The number of positively labeled classes, which falsely have been classified as Negative "incorrectly predicts the Negative class".

In this work, we have been performed two different experiments to evaluate our proposed classifier on the Cleveland Heart Disease dataset. Where the dataset is divided into the training set and testing set 65%:35% respectively. In the training set, 197 samples of the dataset are used, and 106 samples are examined in the testing set. The first experiment is applied to predict with 13 attributes before using the Naive Bayes feature selection step, while the second experiment is applied on the best subset of 6 attributes identified by the Naive Bayes method.

In the first experiment, the different algorithms (SVM, RF, DT, and KNN) were applied as classifiers to make a prediction on the heart disease dataset with 13 features before using Naive Bayes feature selection. The classification accuracy obtained is shown in Table 2.

Table 2: Proposed System Performance for prediction with 13 features

Algorithm used	True Positive	True Negative	False Positive	False Negative	Accuracy	Sensitivity	Specificity
KNN	52	40	9	5	87%	91%	82%
DT	49	39	10	8	83%	85%	79%
RF	50	40	9	7	85%	87%	81%
SVM	46	41	8	11	82%	80%	83%

In the second experiment, the Naive Bayes algorithm use in the feature selection step before the classification phase to select the best subset of features for the classification process. Accordingly, this approach has selected 6 features as a new subset in an attempt to decrease the classification error and to determine the best accurate result. Hence, applying all the proposed algorithms (SVM, DT, RF, and KNN) on the new subset of features to get a prediction. The experimental observation shows that the proposed system can diagnose heart disease with 98% accuracy with the assist of the SVM. Moreover, it achieved 97%, 95%, 93% of accuracy when using DT, KNN, RF classifiers respectively.

The selection of attributes is critically important because including irrelevant attributes in our dataset used for training can result in overfitting. For example, Decision tree algorithms seek to make optimal splits in attribute values. Those attributes that are more correlated with the prediction are split on first. So, when all features are used to train the tree to make a decision, the less relevant and irrelevant attributes will be used to make a prediction decision and this leads to an overfitting problem. This overfitting of the training data can negatively affect the modeling power of the algorithm and cripple the predictive accuracy. It is important to remove redundant and irrelevant attributes from our training dataset before applying algorithms.

Dependence measures qualify the ability to predict the value of one variable from the value of another. The conditional probability is a classical dependence measure that and can be used to find the correlation between a feature and a class. If the correlation of feature X with class C is higher than the correlation of feature Y with C, then feature X is preferred to Y. This procedure is used to determine the dependence of a feature on other features. In our study, the Naive Bayes algorithm was implemented as one of the dependence measures to increase the accuracy of system performance through reduces overfitting and avoid misleading data. As shown in Table 3 the performance of classifiers increases as the number of features decreases.

Table 3: Proposed System Performance for prediction with 6 features.

Algorithm used	True Positive	True Negative	False Positive	False Negative	Accuracy	Sensitivity	Specificity
KNN	54	47	2	3	95%	94.7%	96%
DT	55	48	1	2	97%	96.4%	98%
RF	51	45	4	6	93%	89%	92%
SVM	56	48	1	1	98%	98.2%	98%

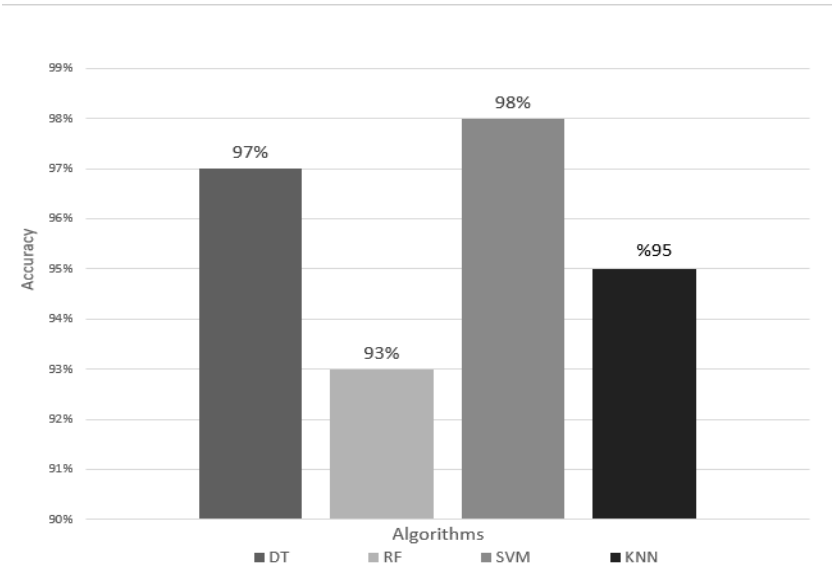

Fig.5: The proposed system (NB-SKDR) performance in term of classification accuracy.

Based on the previous experiment, Fig.5 shows that the improvement is done to increase accuracy and efficiency for a heart disease prediction system by combines several classification techniques with the Naive Bayes approach. The results confirm that accurate prediction can be taken by mixed a Naïve Bayes and SVM which gives better accuracy than other classification techniques.

4.1 Discussion

The proposed study intended to examine the effect of using the Naive Bayes feature selection approach on the accuracy of several classification algorithms in the heart disease prediction system. To verify the performance of the proposed (NDRV-NB) based on Naive Bayes as a feature selection technique, it has been compared with two other Heart Disease Prediction systems. Where the first comparing approach that proposed by [7] was implemented by using the GA (Genetic Algorithm) as a feature selection method and the SVM RBF algorithm as a classifier. While in the second comparative study conducted by [8], the prediction system was implemented by using the PCA (Principal Component Analysis) as a feature selection technique and SVM RBF as a classifier. The comparative performance outcomes are shown in Table (4).

Table 4: A comparative performance studies in term of classification accuracy of the proposed system based on Naïve Bayes feature selection and other system based on PCA and GA feature selection.

The Systems	Database	Selection Approach	Classifier	Accuracy
(Iftikhar *et al.*, 2017)	Cleveland Heart Disease dataset	GA	SVM RBF	88.1%
(Sai Santosh *et al.*, 2019)	Cleveland Heart Disease dataset	PCA	SVM RBF	54.3%
The proposed system	Cleveland Heart Disease dataset	NB	SVM RBF	98%

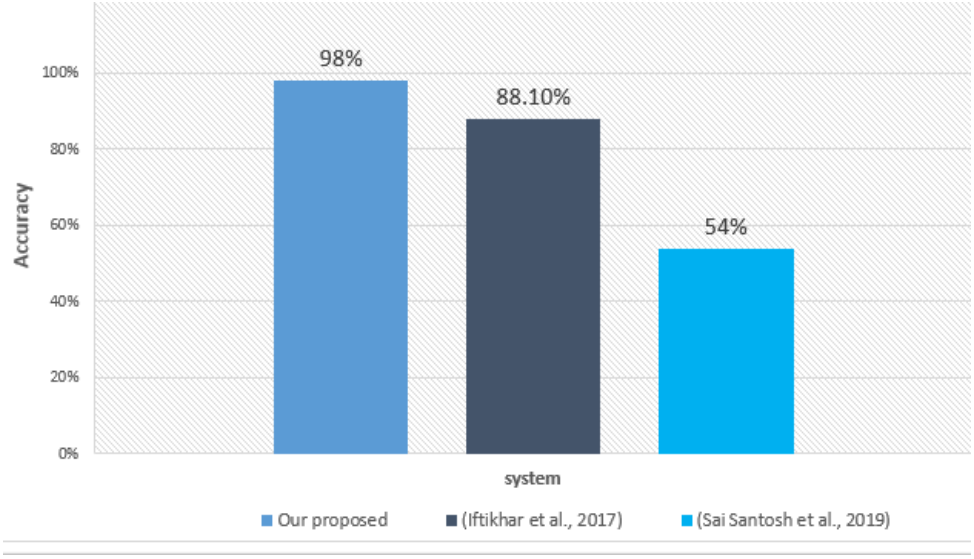

Fig.6: Comparative performance results of accuracy between the proposed approach and other approaches on the same Heart Disease Dataset.

From Table (5) and Fig (6) noticed that the proposed system based on the Naive Bayes technique superior on the other classifiers, And give better results than the other classifiers in terms of classification accuracy compared with other prediction systems introduced by [8], and [7] which are using GA and PCA method in feature selection step .the proposed result was 98% of accuracy, Whereas the other systems based on PCA and GA produced 54%, 91.3% of accuracy respectively. And thus, the effectiveness of the proposed prediction system of heart disease was validated by comparison with other prediction systems.

5. Conclusion

The main aim of this thesis is to propose a new mixed heart disease prediction model built on the Naive Bayes approach with different classification algorithms and based on the Cleveland Heart Disease dataset. The proposed system contains three major stages which include: Preprocessing, Feature Selection, and Classification. Whereas the main achievement of this study was to improve the performance of the heart disease prediction system and to discover the best subset of features that achieve high efficiency of the classification process.
First, the preprocessing stage prepared the data for the next stage by applying two operations which are: filtering method, and discretization. At the preprocessing operations, the null data are firstly handled by replacing with a mean value that was evaluated from every attribute in which the data was null. After that, the data was discretized into a new specific range in an attempt to make the dataset well-structured and fit for the next phase (Feature Selection).
In the feature selection stage, (NB) method based on Bayes Theorem has been adopted as a reduction algorithm to reduce the high dimensionality of the features and select the best subset of dependent features. The (NB) method is used to find the most important features by calculating the dependent probability between each pair of features using the Bayes rule. This method has been able to reduce features from 13 to 6 which are (age, gender, blood pressure, fasting blood sugar, cholesterol, exercise induce engine). The classification stage used several supervised machine learning algorithms as classifiers including (SVM, RF, DT, and KNN) to predict heart disease. From the study, it has been concluded that SVM achieved the highest accuracy of 98% for heart disease prediction. The proposed method was compared with other methods based on PCA and GA, it was found that the (NB) based on Bayes theorem is a promising approach for feature selection in terms of a classification accuracy rate and the number of selected features.

References

[1] Brendan, M., & Reilly, M. D. (2018). The Best Medical Care in the World. *The new England Journals of medicine*, pp. 684–688.
[2] Yang, J. J. et al. (2015). Emerging information technologies for enhanced healthcare. Computers in Industry. vol. 69, pp. 3–11.
[3] Jamse et al. (2018). Design and Implementation of a Hospital Database Management System (HDMS) for Medical Doctors. International Journal of Computer Theory and Engineering, 10(1), pp.1–6.
[4] Razeghi, R., & Nasiripour, A. A. (2014). An investigation of factors affecting Electronic
[5] Rajkumar, A., & Reena, G. (2010). Diagnosis of heart disease using datamining algorithm. Global journal of computer science and technology, 10(10), pp. 38–43.
[6] Vaddella, D.,Sruthi, C., Chowdary, B., Subbareddy, R., & Somula, G. (2019). Prediction of heart disease using machine learning techniques. *International Journal of Recent Technology and Engineering*, 8(2 Special Issue 4), pp. 612–616.
[7] Iftikhar, S., Fatima, K., Rehman, A., Almazyad, A. S., & Saba, T.(2017). An evolution based hybrid approach for heart diseases classification and associated risk factors identification, *Biomedical Research (India)*, 28(8), 3451–3455.
[8] Santosh, B., Reddy, D., Vardhan, M., & Subhani, S. (2019). Heart Disease Prediction with PCA and SVM, *International Journal of Engineering and Advanced Technology (IJEAT)*, (4), pp. 2249–8958.
[9] Kaur, G., Sharma, Anshu, and Sharma, Anurag. (2019). Heart Disease Prediction using KNN classification approach. *international Journal of Computer Sciences and Engineering*, 7(5) ,pp. 416–420.
[10] Ghorbani, R. & Ghousi, R. (2019). Predictive data mining approaches in medical diagnosis: A review of some diseases prediction. International Journal of Data and Network Science, vol. 3, pp. 47–70.
[11] Shukla, N., & Arora, M. (2016). Prediction of diabetes using neural network & random forest tree. International Journal of Computer Sciences and Engineering, vol. 4, pp. 101-104.
[12] Pouriyeh, S., Vahid, S., Sannion, G., Pietro, G. D., Arabnia, H., & Gutierrez, J. (2017). A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. Proceedings-IEEE Symposium on Computers and Communications, (Iscc), pp. 204–207.
[13] Abdar, M., Kalhori, S. R., Sutikno, T., Ibnu Subroto, I. M., & Arji, G. (2015). Comparing performance of data mining algorithms in prediction heart diseses. International Journal of Electrical and Computer Engineering, 5(6), pp. 1569–1576.
[14] Ming, D., Wang, S. M., & Gong, G. (2011). Research on decision tree algorithm based on information entropy, Advanced Materials Research, vol. 267.
[15] Nowozin, S. (2012). Improved information gain estimates for decision tree induction, Proceedings of the 29th International Conference on Machine Learning, ICML 2012, 1, pp. 297–304.

[16] Cotrtes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), pp. 273-297.
[17] Maji, S., Berg, A., & Malik, J. (2008). Classification using intersection kernel support vector machines is efficient. IEEE conference on computer vision and pattern recognition, pp. 1-8. IEEE.
[18] Patle, A., & Chouhan, D. S. (2013). SVM kernel functions for classification, 2013 International Conference on Advances in Technology and Engineering, ICATE 2013.
[19] Shiliang, S., & Rongqing, H. (2010). An adaptive k-nearest neighbor algorithm. In 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, vol. 1, pp. 91–94.
[20] Abu Alfeilat, H. A. et al. (2019). Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. Big Data, 7(4), pp. 221–248.
[21] Wiener, A., & Liaw, M. (2003). Classification and Regression by random Forest. International Journal of Innovative Research in Science, Engineering and Technology, pp. 18–22.
[22] Ho, T. K. (1998). The Random Subspace Method for Constructing Decision Forests, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 20(8), pp. 832–844.
[23] Zhang, H. (2005). Exploring conditions for the optimality of naïve bayes, *International Journal of Pattern Recognition and Artificial Intelligence*, 19(2), pp. 183–198.
[24] Alasadi, S. A. and Bhaya, W. S. (2017) 'Review of data preprocessing techniques in data mining', Journal of Engineering and Applied Sciences, 12(16), pp. 4102–4107. doi: 10.3923/jeasci.2017.4102.4107.
[25] Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World, Annual Review of Psychology, 60(1), pp. 549–576.
[26] Abraham, R., Simha, J. B., & Iyengar, S. S. (2006). A comparative analysis of discretization methods for medical datamining with Naïve Bayesian classifie. Proceedings-9th International Conference on Information Technology, ICIT 2006, pp. 235–236.
[27] Purpura, A., Masiero, C., Silvello, G., & Susto, G. (2019). Feature selection for emotion classification, CEUR Workshop Proceedings, vol. 2441, pp. 47–48.
[28] Xue, B., Zhang, M., & Browne, W. N. (2012). Particle Swarm Optimization for Feature Selection in Classification : A Multi-Objective Approach, IEEE Transactions on Cybernetics, pp. 1–16.
[29] Marlina, L., lim, M. & Siahaan, A. P. (2016). Data Mining Classification Comparison (Naïve Bayes and C4.5 Algorithms), International Journal of Engineering Trends and Technology, 38(7), pp. 380–383.
[30] Konieczny, R. & Idczak, R. (2016). Supervised Machine Learning: A Review of Classification Techniques, Hyperfine Interactions, 237(1), pp. 1–8.