The Role of Data Pre-processing Techniques in Improving Machine Learning Accuracy for Predicting Coronary Heart Disease

Osamah Sami[1], Yousef Elsheikh[2], Fadi Almasalha[3]
Faculty of Information Technology
Applied Science Private University
Amman, Jordan 11931

***Abstract*—These days, in light of the rapid developments, people work day and night to live at a good level. This often causes them to not pay much attention to a healthy lifestyle, such as what they eat or even what physical activities they do. These people are often the most likely to suffer from coronary heart disease. The heart is a small organ responsible for pumping oxygen-rich blood to the rest of the human body through the coronary arteries. Accordingly, any blockage or narrowing in one of these coronary arteries may cause blood not to be pumped to the heart and from it to the rest of the body, and thus cause what is known as heart attacks. From here, the importance of early prediction of coronary heart disease has emerged, as it can help these people change their lifestyle and eating habits to become healthier and thus prevent coronary heart disease and avoid death. This paper improve the accuracy of machine learning techniques in predicting coronary heart disease using data preprocessing techniques. Data preprocessing is a technique used to improve the efficiency of a machine learning model by improving the quality of the feature. The popular Framingham Heart Study dataset was used for validation purposes. The results of the research paper indicate that the use of data preprocessing techniques had a role in improving the predictive accuracy of poorly efficient classifiers, and shows satisfactory performance in determining the risk of coronary heart disease. For example, the Decision Tree classifier led to a predictive accuracy of coronary heart disease of 91.39% with an increase of 1.39% over the previous work, the Random Forest classifier led to a predictive accuracy of 92.80% with an increase of 2.7% over the previous work, the K-Nearest Neighbor classifier led to a predictive accuracy of 92.68% with an increase of 2.58% over the previous work, the Multilayer Perceptron Neural Network (MLP) classifier led to a predictive accuracy of 92.64% with an increase of 2.64% over the previous work, and the Naïve Bayes classifier led to a predictive accuracy of 90.56% with an increase of 0.66% over the previous work.**

Keywords*—Coronary heart disease; heart; machine learning; data preprocessing; classification technique

I. Introduction

The heart is one of the most important organs in the human body. It is a small, muscular pumping organ responsible for supplying other organs in the body with oxygen and other important nutrients [1]. This means that a person's life depends on the efficiency of heart function. Therefore, if the heart does not function well, other organs also cannot function well [2].

People, in light of the difficult economic conditions, seek to secure their basic needs by working long hours daily. This lifestyle often does not take into account the diet and health of these people to ensure their safety [3]. This type often leads to a risk of diseases such as diabetes, high cholesterol and blood pressure at an early age, and all of these diseases, if not controlled, can lead to coronary heart disease [3].

Heart disease is a term that refers to any problem that can affect the heart and blood vessels [2], such as coronary heart disease, congenital heart disease, and rheumatic heart disease [4], which, according to the National Heart, Lung, and Blood Institute ranks among the most dangerous and common diseases in the world.

In coronary heart disease, a complete or partial blockage of the coronary arteries usually occurs due to blood clotting or the accumulation of fatty plaques on the walls, which leads to the inability of the heart to get enough oxygen [5] and thus it is difficult for the heart to function as efficiently as required.

There are two risk factors for coronary heart disease. The first type is stable and cannot be changed, such as age, gender and family history, while the other type depends on lifestyle such as diabetes, smoking, high cholesterol, high blood pressure, high body mass index, and low exercise [6]. However, the second type of risk factors can usually be controlled, according to experts, by changing our lifestyle and diet, and using certain medications if needed.

In recent years, artificial intelligence techniques have been used extensively in the medical fields in order to improve the efficiency of disease diagnosis/classification in its early stages [7]. Among those techniques stand out machine learning techniques, which are a set of statistical models that help the machine learn from past data [8]. In spite of this, it is often difficult to deal with patient data for diagnosis in the early stages due to reasons such as data volume, missing values and noise in the data. But machine learning techniques and their capabilities have helped process such data [9].

Also, it is noticeable regarding data features that they may be incomplete and huge. The range of some data features is small while the range is large for other data features. The type of data features is combined between categorical and numerical; all of this will affect the accuracy of machine learning techniques in diagnosing and classifying diseases in their early stages, including coronary heart disease. Using different techniques to manipulate the features under the so-called data preprocessing techniques and thus improve the

accuracy of machine learning techniques in early prediction of the disease [10]. 12 This 6earc21paper is organized as follows: The second section is a review of some relevant work. The third section presents the methodology for this research paper. The fourth section is for presenting, evaluating and discussing the results of the research paper. The fifth section is for conclusion and the sixth section is the future work.

II. Related Work

Recently, there has been an increase in the number of papers dealing with the use of machine learning techniques in predicting serious diseases that may affect people's lives, including coronary heart disease. In [11], the researchers applied a logistic regression technique on the Framingham Heart Study dataset to predict the ten-year risk of coronary heart disease. The researchers used 65% of the dataset for the training set. The accuracy obtained was 84.8%.

The researchers in [12] had a contribution by implementing four machine learning algorithms, namely support vector machine (SVM), neural network, XGBoost, and random forest to predict the ten-year risk of coronary heart disease. The researchers also used the Framingham Heart Study dataset to validate the results. The accuracy obtained was 84.8% for support vector machine, 85.4% for neural network, 86.99% for XGBoost, and 84.9% for random forest.

Also, the researchers in [4] contributed to the literature of this field by using boosting adaptive algorithm on four datasets, namely (UCI Cleveland, UCI Switzerland, UCI Long Beach, and UCI Hungarian) to diagnose coronary heart disease. This approach obtained accuracy (97.16% and 80.14% for Cleveland, 98.63% and 89.12% for Hungarian, 93.15% and 77.78% for Long Beach, 100% and 96.72% for Switzerland) for training and testing set respectively.

In [13], the researchers applied three machine learning algorithms, namely support vector machine, neural network, and Hybrid-SVM on the Framingham Heart Study dataset to predict the ten-year risk of heart attack. The accuracy obtained was 86.03% for support vector machine, 84.7% for neural network, and 94% for Hybrid-SVM. However, these results were better for some of the machine learning techniques used than those used for [12].

In [14], the researchers applied six algorithms, namely decision tree, boosted decision tree, random forest, support vector machine, neural network, and logistic regression on the Framingham Heart Study dataset to predict the ten-year risk of coronary heart disease. The data was divided into 80% training and 20% testing. The researchers used R Studio and Rapid-Miner in their work. The researchers used three techniques to deal with missing values. The first technique is to ignore missing values, and obtained accuracy of 85% for the decision tree, 63% for the boosted decision tree, and 63% for logistic regression. All this while using the Rapid-Miner tool. Whereas, the R studio tool enabled the researchers to obtain the accuracy of 84% for the decision Tree, 85% for the boosted decision tree, and 84% for logistic regression. Analysis of complete case is the second technique used, as the Rapid-Miner tool enabled the researchers to obtain accuracy of 54% for the decision tree, 64% for the boosted decision tree, 65% for the random forest, 69% for the support vector machine, 69% for the neural network, and 68% for logistic regression. R studio tool obtained accuracy 67%, 81%, 79%, 69%, 67%, and 68% for the decision tree, boosted decision tree, random forest, support vector machine, neural network, and logistic regression respectively. The final technique is to be replaced with the average, and the accuracy obtained while using the Rapid-Miner tool was 62% for the decision tree, 62% for the boosted decision tree, 63% for the random forest, 68% for the support vector machine, 68% for the neural network, and 67% for logistic regression. Whereas, the R Studio tool enabled the researchers to obtain an accuracy of 84% for the decision tree, 84% for the boosted decision tree, 78% for the random forest, 68% for the support vector machine,71% for the neural network, and 66% for logistic regression.

However, other researchers such as those in [15] applied only one algorithm which is the logistic regression on the Framingham Heart Study dataset to predict the ten- year risk of coronary heart disease. This approach obtained better accuracy of 86.6% than ever.

In [16], the researchers applied the same previous method of logistic regression to the Framingham Heart Study dataset to predict a heart attack. This approach obtained an accuracy of 87%.

Other researchers such as those in [17] applied the neural network algorithm to real data from patient of Paris Hôtel-Dieu University Hospital to diagnose coronary heart disease. Their approach used a different number of input factors (6 to 14). The approach obtained 63% for features (age, diabetes, hypertension, obesity, smoking, family anamnesis of CHD), 76% for features (age, sex, diabetes, hypertension, obesity, smoking, family anamnesis of CHD), 77% for features (age, sex, diabetes, hypertension, obesity, smoking, family anamnesis of CHD, glycaemia, cholesterol total), 81% for features(age, sex, diabetes, hypertension, obesity, smoking, family anamnesis of CHD, TG, cholesterol 0.81 69 79 total, HDL, LDL, glycaemia), 83% for features (age, sex, diabetes, hypertension, obesity, smoking, family anamnesis of CHD, carotid plaque), 87% for features (diabetes, hypertension, obesity, smoking, family anamnesis of CHD, PWV index), 91% for features (diabetes, hypertension, obesity, smoking, family anamnesis of CHD, carotid plaque, PWV index), 93% for features (diabetes, hypertension, obesity, smoking, family anamnesis of CHD, TG, cholesterol, HDL, 0.93 80 92 LDL, glycaemia, carotid plaque, PWV index), 77% for features (age, sex, diabetes, hypertension, obesity, smoking, family anamnesis of CHD, glycaemia, 0.77 53 87 cholesterol total, cGFR), and 77% for features (age, sex, diabetes, hypertension, obesity, smoking, family anamnesis of CHD, glycaemia, cholesterol total, left ventricular hypertrophy)

Those in [18] applied the deep belief algorithm to the KNHANES-6 dataset to predict the risk of coronary heart disease and obtained an accuracy of 82%. However, the researchers applied the genetic algorithm to improve the deep belief network and the obtained accuracy was 74%.

In [19], the researchers applied a logistic regression and neural network to the KNHANES-VI dataset to predict the risk of coronary heart disease. However, this approach obtained accuracy 86.11% for the logistic regression and 87.04% for the neural network. The researchers used a distinct correlation

analysis to improve the accuracy of the neural network to become 87.63%.

In other research such as [20], the researchers applied Naïve Bayes, KNN, random forest, decision tree, SVM, logistic regression, and the ensemble classification approach to the NHANES and Framingham Heart Study dataset, to monitor the risk of chronic diseases. For the NHANES dataset, the decision tree algorithm obtained an accuracy of 97.6%, 96.5% for the ensemble approach, 80.8% for the KNN, 96.4% for logistic regression, 95.7% for Naïve Bayes, 98.5% for random forest, 95.4% for SVM. Whereas, the results for Framingham Heart Study dataset were as follows: The decision tree obtained an accuracy of 90%, 89.3% for the ensemble approach, 90.1% for the KNN, 90% for the logistic regression, 89.9% for Naïve Bayes, 90.1% for random forest, and 90.2% for SVM.

Similarly, the researchers of [21] applied Naïve Bayes, KNN, random forest, decision tree, SVM, logistic regression, neural network, and the ensemble classification approach to the NHANES and Framingham Heart Study dataset to predict Cardiovascular disease. For the NHANES dataset, the decision tree algorithm obtained an accuracy of 97.6%, 96.5% for the ensemble approach, 80.8% for the KNN, 96.4% for logistic regression, 95.7% for Naïve Bayes, 98.5% for random forest, 95.4% for SVM, 98.8% neural network. Whereas, the results for Framingham Heart Study dataset were as follows: The decision tree obtained an accuracy of 90 89.3% for the ensemble approach, 90.1% for the KNN, 90% for logistic regression, 89.9% for Naïve Bayes, 90.1% for random forest, 90.2% for SVM, and 89% for neural network.

In [22], the researchers applied neural network algorithm on the Framingham Heart Study dataset to predict the heart disease. The accuracy obtained was 90% .

Other researchers such as those in [23] applied the k-nearest neighbor (KNN), Logistic regression (LR), linear discriminant analysis (LDA), support vector machine (SVM), classification and regression tree (CART), gradient boosting (GB), and random forest (RF) the Framingham Heart Study dataset to detect the heart disease. The accuracy obtained was 81% for KNN, 83% for LR, 83% for LDA, 82% for SVM, 75% for CART, 83% for GB, and 83% for RF. After that some ensemble techniques were applied and the accuracy was improvement to 86%.

Those in [24] applied k-nearest neighbor, decision tree, random forest logistic regression, and neural network on the Framingham Heart Study dataset to predict the heart disease. The accuracy obtained was 86% for k-nearest neighbor, 77% for decision tree, 86% for random forest, 85% for logistic regression, and 85% for neural network.

Most of previous researchers using either the UCI dataset or Framingham Heart Study dataset, UCI dataset is a good dataset for diagnosis, and prediction heart disease, but this data has some limitations, first limitation is the size of instance of the data is bit small, second limitation the dataset does not include some important features for predict and diagnose heart disease such as LDL cholesterol, HDL cholesterol, smoking or not smoking, diastolic blood pressure, systolic blood presume, number of cigarettes per day, body mass index, and family history of any type of heart disease. This means this data does not fit to diagnose or predict heart disease for smoking patients, patient with history of blood pressure, obesity patients, and patients with a family history of heart disease.
also, Framingham Heart Study dataset is good data for predict heart disease, this data does not contain feature for family history of any type of heart disease. This means this data specific for patient with no family history of any type of heart disease.

Despite this and many other researches, the field is still open for researchers to conduct their experiments in order to improve the accuracy of the machine learning techniques for predicting diseases that pose a risk to human life, including coronary heart disease.

III. Research Method

It is unfortunate to hear that there is an increase in the number of patients diagnosed with coronary heart disease (angina or heart attack) day after day. High blood pressure, high cholesterol, uncontrolled diabetes, smoking, and a diagnosis of cardiovascular impairment and other risks, all increase the chance of diagnosis with coronary heart disease in the future. Therefore, an accurate system needed to help the patient protect him/herself from the risk of coronary heart disease, relying in this on the patient's demographic information, medical history, medical examination, behavior, and laboratory examination.

Many researchers have developed machine learning models using different classification algorithms such as decision tree, Naïve Bayes, SVM, KNN, and neural network. Most of these models were utilizing the Cleveland Heart Diseases dataset to predict coronary heart diseases, but few were using the Framingham Study dataset. This paper uses the Framingham Study dataset to validate the resulting model since it includes features for most of the potential risk factors for coronary heart disease and some of these features are not found in the most common dataset of heart disease namely, Cleveland Heart Disease dataset. In this paper, five machine learning classification algorithms were used such as decision tree, Naïve Bayes, neural network, random forest, and KNN. These five algorithms used the Framingham Heart Study dataset with two events for target (output) features to predict coronary heart disease, as a number of different Data Preprocessing techniques will be used to improve the accuracy of machine learning models for predicting coronary heart disease.

A. Dataset

The Framingham Heart Study dataset is the first long-term epidemiological study concerned with the possible causes of cardiovascular disease that began in 1948 in Framingham, Massachusetts [20]. The Framingham Heart Study dataset identified the prospective risk factors of cardiovascular diseases and their effects [20], [25].

The dataset contains 19 input features divided into demographic features(Age, Gender), behavioral features(Current Smoker, Cigarettes Per Day, Body Mass Index), medical history features(Prevalent Coronary Heart Diseases, Prevalent Angina Pectoris, Prevalent Myocardial Infarction, Prevalent Stroke, Prevalent Hypertensive, Use Blood Pressure Drugs,

Diabetes), medical examination features(Systolic Blood Pressure, Diastolic Blood Pressure, Heart Rate) and laboratory testing features(Glucose, High-Density Lipoprotein, Low-Density Lipoprotein, Total Cholesterol), and two features for prediction (Angina Pectoris, Myocardial Infarction).

B. Data Preprocessing

Data preprocessing is a group of techniques that are applied on the data to improve the quality of the data, such as handling missing values, convert the type of feature and many other techniques [10].

1) Impute Missing Values By Knn: knn for missing values working by calculate the distance or similarity to find the most similar case in the dataset and change the missing value with it [26], by applying (1).

$$Dist(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \quad (1)$$

Where X_i some known values, and Y_i some values that should predict their values.

2) Min_Max Normalization: This method is convert each numerical feature value into new value depending on the minimum and maximum values of the feature [27], by applying (2).

$$\bar{X} = \frac{X - Min}{Max - Min} \quad (2)$$

Where Min is the smallest value in the selected feature, Max is the biggest value in the selected feature, $\bar{X}$ is a new select value after applying normalization, X is a selected value from a numerical feature.

3) Z-Score Standardization: This method is convert each numerical feature value into new value depending on the standard deviation and Mean of the feature [28], by applying (3).

$$\bar{X} = \frac{X - \mu}{\sigma} \quad (3)$$

$\bar{X}$is a new select value after applying standardization, X is a selected value from a numerical feature.

4) One Hot Encoding: One Hot Encoding splits the categorical feature into a separate number of features depending on the number of the cases in the original categorical feature, and give 0 for absence and 1 for presence in each new feature [29].

5) Ordinal Encoding: In this technique, each case in the categorical feature is converted into integer value [29].

6) Equal Width Discretization: This is an easy method that sorting the values of numerical feature and split the range of sorting values into predefined equal-width bins [30] by applying (4) and (5).

$$W = \frac{V_Max - V_Min}{K} \quad (4)$$

$$Boundaries = V_Min + (i * W) \quad (5)$$

Where W is the width of the bin, V_Max is the maximum value in the selected numerical feature, V_Min is the minimum in the selected numerical feature, i = 1..... k-1.

7) Equal Frequency Discretization: In this method, firstly sorting the values in ascending order. Split the range of sorting values into predefined number of equal-frequency bins by applying $\frac{N}{K}$, each bin has the same number of values [30].

C. Classification Algorithms

Classification is a supervised machine learning model used with a label's output to determine the result of the model from many labels or categorical input data [31]. The classifier model is built for training depending on many known labelled or categorical feature of input data [31]. In the next step, the model tested by using the test set to identify the number of the known target for the model and try to correct the unknown target for the model [31].

1) ID3 Decision Tree: Each decision tree contains a root node, leaf node, internal node and branches. In ID3 decision tree, all features set as root node, and after that the features are divided by finding the entropy which it utilizes the measure of the harmony in the data; the values of entropy is between 0 and 1 [7], and information gain is the difference between the feature and the subsets of this feature [7]. Entropy and information gain can be found by applying (6) and (7), and the feature which has the highest information gain value is selected as the root node of the tree [7].

$$Entropy(F) = \sum_{i=1}^{C}(-P_i \log_2 P_i) \quad (6)$$

$$Gain(F,A) = Entropy(F) - \sum_{i=1}^{K}(\frac{|F_i|}{|F|}Entropy(F_i) \quad (7)$$

Where C is number of outputs, P_i is probability of occurrences each output from all output, K number of spilt data, F feature with some data, F_i spilt data from feature F.

2) Random Forest: Random forest is a classification algorithm [32] works by creating many decision trees from the dataset [32]. The features are selected randomly from the training set to build the trees in the random forest [32]. After building each decision tree and find the result of each the tree, applying majority voting to decide the final result of the random forest [32]. In the process of building each decision tree, the randomization is applied to find the value the split node.

3) K-Nearest Neighbours: KNN is a lazy supervised machine learning algorithm that used to predict and classify unknown data from known data by measuring the distance between them [33]. The distance metric is using to measure the distance between point from testing data with all the point in training data [33], [34], the distance can calculate by applying (8).

$$Cosine(X_i, Y_i) = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n}(X_i)^2}\sqrt{\sum_{j=1}^{n}(Y_i)^2}} \quad (8)$$

Where X_i some values belong to known output class , and Y_i some values that should predict their output class.

4) Multilayer Perceptron Neural Network: Artificial Neural network structure is the same as the brain of human [35]. Multilayer perceptron (MLP) that contains more than one layer**(input layer, hidden layer(s), output layer)** [36].

First, in the neural network before start training from the dataset, the value of weight (w) is randomly assigned [36]. After that, the neural network begin the training [36]. Sigmoid is a non-linear activation function commonly use in feedforward neural networks to find the output [37]. Sigmoid function can be calculated by applying (9).

$$F(X) = \frac{1}{1 + \exp^{-X}} \tag{9}$$

Back Propagation algorithm is commonly used to train Multilayer Perceptron Neural Network In the first step of this algorithm is to compare between predict output $(\bar{Y})$ and actual output **(Y)** to find the error between them, this error return to neural network and the weight change depending on this error, and the weight numerical change until the value of$(\bar{Y})$ become closer to **(Y)** [36].

5) Naïve Bayes: Naïve Bayes is a statistical classification algorithm that works on the basis of Bayes' theory, and Naïve Bayes assumes that each feature is separate, and each variable is distinct in prediction and occurrence [3]. Naïve Bayes uses the prior probability of Bayes theorem to calculate the likelihood of the relationship between each feature in the test data with each target, the target with the highest probability is selected as the result of the model [38]. The probability can be found using (10):

$$P(C_i|F_j) = \frac{P(F_j|C_i)P(C_i)}{P(F_j)} \tag{10}$$

Where $P(C_i|F_j)$ probability of specific class (C_i) appear with specific feature (F_j) from the total of all Features F and Classes C, $P(C_i)$ probability of specific class (C_i) from the total of all classes (C), $P(F_j|C_i)$ probability of specific feature (F_j) appear with specific class (Ci) from the total of all features (F) and classes(C), $p(F_j)$ probability of specific feature (F_j) from the total of all features (F).

D. Stratified KFold Cross Validation

Cross validation is a static method used to test an algorithm by dividing the data set into a training set used to train the model and the test set used to evaluate the model performance [39]. In cross-validation, every point has the same chance of being used in the test [39]. In kfold, the dataset is evenly divided into k number of fields [39]. Stratified KFold means that each fold has the same class naming distribution in the original dataset [40]. For each iteration, one test folds and others are used for training [39].

E. Tool

RapidMiner is a data science software platform developed by the company of the same name that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics [41]. In machine learning, RapidMiner can be used for feature processing, dataset segmentation, model training, model testing, network research, and performance evaluation [41].

IV. Results and Discussion

In this paper, five machine learning classification techniques used to predict two primary CHD events, namely, angina pectoris (528 yes, 2735 no) and myocardial infarction (308 yes, 2955 no).

A. Performance Evaluation

Performance evaluation is a group of equations used to measure the effectiveness of the classifier or the model [42]. Below is the definition of some essential terms used in the equations of performance evaluation:

1) True Positive (TP): The person is healthy and also predict as healthy [42]

2) False Positive (FP): The person is healthy, but predict as sick [42]

3) True Negative (TN): The person is sick and predict as sick [42]

4) False Negative (FN): The person is sick, but predict as healthy [42]

B. Confusion Matrix

The confusion matrix is used to analyze the ability of classifier or model to identify the classes of the dataset [42]. TN and TP are referred to correct classification, while FN and FP are referred to wrong classification [42]. For the accurate classifier or model, TP and TN are classified more than FN and FP [42], as shown in Table I

TABLE I. Confusion Matrix

	Negative(Actual)	Positive(Actual)
Negative(Predict)	TN	FN
Positive(Predict)	FP	TP

C. Performance Metrics

1) Accuracy: Accuracy is an evaluation metric of the total number of predictions the model or the classifier gets right [43]. The accuracy can be calculated by applying (11).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{11}$$

2) Precision: Precision is used to identified is the diagnosis or the predicted result is close to the real result [43]. Precision can be calculated by apply (12).

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

3) F-Measure: F-Measure refers to the mean of consistency between Precision and Recall [43]. F-Measure can be calculated by apply (13).

$$F - Measure = 2 * (\frac{Recall * Precision}{Recall + Precision}) \tag{13}$$

4) Sensitivity(Recall): Sensitivity is true positive rate measure. In other words, the rate of healthy person diagnosis or predict as healthy [43]. Sensitivity can be calculated by apply (14).

$$Sensitivity = \frac{TP}{TP + FN} \tag{14}$$

5) Specificity: Specificity is true negative rate measure. In other words, the rate of sick person diagnosis or predict as sick [43]. Specificity can be calculated by apply (15).

$$Specificity = \frac{TN}{TN + FP} \tag{15}$$

D. Algorithms Confusion Matrix

Below Table II, Table III, Table IV, Table V, and Table VI, shown the number of correct predict **(True Positive and True Negative)** and wrong predict **(False Positive and False Negative)** for each algorithm.

TABLE II. ID3 DECISION TREE CONFUSION MATRIX

	No(Actual)	Yes(Actual)
No(Predict)	2879	205
Yes(Predict)	76	103

TABLE III. RANDOM FOREST CONFUSION MATRIX

	No(Actual)	Yes(Actual)
No(Predict)	2921	201
Yes(Predict)	107	34

TABLE IV. K-NEAREST NEIGHBORS CONFUSION MATRIX

	No(Actual)	Yes(Actual)
No(Predict)	2930	214
Yes(Predict)	94	25

TABLE V. NEURAL NETWORK CONFUSION MATRIX

	No(Actual)	Yes(Actual)
No(Predict)	2923	208
Yes(Predict)	100	32

TABLE VI. NAÏVE BAYES CONFUSION MATRIX

	No(Actual)	Yes(Actual)
No(Predict)	2666	239
Yes(Predict)	69	289

E. Accuracy without Data Preprocessing

TABLE VII. ACCURACY WITHOUT DATA PREPROCESSING

Algorithms	Accuracy (%)
Decision Tree	87.19
Random Forest	92.68
MLP	90.56
KNN	90.50
Naïve Bayes	89

F. Algorithms Evaluation Result

TABLE VIII. MODEL EVALUATION RESULT

Algorithms	Accuracy	Precision	F-Measure	Sensitivity	Specificity
ID3 Decision Tree	92.8%0	93.57%	96.13%	98.85%	34.80%
Random Forest	91.39%	93.36%	95.35%	97.43%	33.50%
K-Nearest Neighbors	92.68%	93.20%	96.08%	99.15%	30.53%
Neural Network	92.64%	93.36%	96.06%	98.92%	32.51%
Naïve Bayes	90.56%	91.79%	94.54%	97.48%	54.77%

G. Accuracy Comparison

TABLE IX. ACCURACY COMPARISON

Algorithms	Previous Accuracy	Proposed Accuracy	Dataset Event
Decision Tree	90% [20], [21]	91.39%	Myocardial Infraction
Random Forest	90.1% [20], [21]	92.80%	Myocardial Infraction
K-Nearest Neighbors	90.1%[20], [21]	92.68%	Myocardial Infraction
Neural Network	90% [22]	92.64%	Myocardial Infraction
Naïve Bayes	89.9% [20], [21]	90.56%	Angina Pectoris

H. Discussion

In this research paper, a set of machine learning techniques used to predict two events of coronary heart disease namely, Angina Pectoris (528 Yes, 2735 No), and Myocardial Infarction (308 Yes, 2955 No). Despite the previous researchers used many data preprocessing techniques, the results obtained from this work were very encouraging compared to other studies that use the same data set to calculate accuracy as shown in Table IX.

It is noted that the techniques that have been used to improve the accuracy of machine learning models or classifiers in predicting coronary heart disease have proven effective and thus have achieved better results than previous research.

For example, [20] and [21] used the same data set and obtained by applying the decision tree algorithm a predictive accuracy of 90% to predict coronary heart disease (CHD), while this research paper obtained an accuracy of 91.39%, with a positive increase of 1.39% as shown in Table IX.

Also, this research paper and through the application of the random forest algorithm obtained a predictive accuracy of CHD 92.80%, shown in Table IX, which is higher than the result obtained in the decision tree algorithm in this research paper on the one hand, and on the other hand, higher and better than the results obtained by [20] and [21] and that was 90.10%, with a positive increase of 2.7%.

As for the use of the MLP algorithm in predicting CHD, researchers in [21] obtained an accuracy of predicting the

disease 90%, while this research paper obtained a better accuracy of 92.64%, with a positive increase of 2.64% shown in Table IX.

Regarding the use of the KNN algorithm, researchers in [20] and [21] obtained a prediction accuracy of 90.10%, which is less than the prediction accuracy of the disease obtained in this research paper, which is 92.68%, which was applied to calculate the missing values and equal width discretization, with a positive increase of 2.58% as shown in Table IX.

The application of the Naïve Bayes in this research paper obtained a predictive accuracy of coronary heart disease 90.56% as shown in Table Table IX, which is better than the predictive accuracy of 89.90% obtained in [20].

After applied data preprocessing techniques, this proposed work obtained accuracy better than previous researches used the same dataset and same techniques, such as, [13] that published in **2018** was obtained accuracy 84.7% for neural network; decision tree was obtained 85%, random forest was obtained 79%, and neural network was obtained 71% in [14] that published in **2017**; [20] that published in **2017** was obtained accuracy 90.1% for KNN, 90.1% for random forest, 89.9% for Naïve Bayes, and 90% for decision tree; the accuracy in [21] that published in **2018** was obtained 90.1% for KNN, 90.1% for random forest, 89.9% for Naïve Bayes, and 90% for decision tree; in **2020** the [22] was obtained accuracy 90% for neural network; [23] that published in **2021** was obtained accuracy 81% for KNN, 75% for decision tree, and 83% for random forest; decision tree was obtained 77%, random forest was obtained 86%, KNN was obtained 86%, and neural network was obtained 85% in [24] that was published in **2021**.

Although the results obtained in predicting coronary heart disease in terms of accuracy were not as significant as it should be, it may contribute to an increase in the number of cases with the correct diagnosis of the disease and at the same time reduce the number of cases that are incorrectly diagnosed with coronary heart disease and thus save lives

V. Conclusion

The heart is among the most important organs of the human body, as any problem with it can damage other important organs in the body, such as the brain. All doctors around the world warn of the sharp increase in the number of heart patients, being a serious disease that may lead to serious complications such as heart failure and cardiac arrest, both of which often lead to death if not diagnosed early.

In this paper, the researchers contributed to improving the accuracy of machine learning classification models in predicting two primary coronary heart disease events, namely, angina pectoris and myocardial infarction through the use of a number of feature processing techniques such as normalization, standardization, and discretization. For the purpose of validating the results obtained, the data set of the Framingham Heart Study was used with two main events (angina pectoris and myocardial infarction (heart attack)), due to its containment and after consulting with cardiologists about the most common factors causing coronary heart disease.

After using data preprocessing techniques on the dataset, the accuracy of machine learning algorithms for predicting coronary heart disease improved unevenly. For example, the improvement in accuracy prediction of CHD was 4.2% when using the ID3 decision tree algorithm, 0.14% when using the random forest algorithm, 3.18% when using the KNN algorithm, 2.08% when using the MLP algorithm, and 1.36% when using the Naive Bayes algorithm as shown in Table VII and Table VIII.

However, the best prediction accuracy obtained for the ID3 decision tree algorithm is at 91.39% when applied the equal width discretization method. Whereas, the random forest algorithm achieved a prediction accuracy of 92.80% when applied the equal width discretization and applied normalization methods. The MLP algorithm achieved an improvement in accuracy prediction by 92.64% when using one hot encoding technique. 92.68% represents the predictive accuracy obtained with the KNN algorithm when applied the ordinal coding and standardization techniques. However, all of the predicted values obtained were in the case of a myocardial infarction event. Whereas, the value obtained from Naive Bayes algorithm was 90.65% in the case of angina pectoris and when applied equal frequency discretization. The results obtained confirm the importance of using data preprocessing techniques in improving the accuracy performance of machine learning algorithms for predicting coronary heart disease compared to previous published research with the same objectives.

In the end, the presence of a correlation between some serious diseases such as the occurrence of stroke, high blood pressure, cardiovascular disease and coronary heart disease leads us in the future to predict such diseases and the effect of each of them on the occurrence of coronary heart disease on the one hand, and on the other hand the effect of the occurrence of coronary heart disease, on these diseases, to prevent death. This is because the patient in such cases does not have enough time to go to the doctor to see him and save his life.

VI. Future Work

In the future work, more data preprocessing techniques and more machine learning classification algorithms can apply to get better results than the ones that obtained in this proposed work.

Machine learning algorithms can used to analyze big data to forecast coronary heart disease. This means that a huge amount of data means that the prediction will get better because more data means that the result is more accurate.

Sometimes the patient does not have enough time to go to the doctor, so develop a website or smartphones application for the graphical user interface solve this problem, and this site makes the prediction process easier and from the patient's place where the user only enters his risk factors information and the result is presented to him immediately.

Acknowledgment

The authors are grateful to the Applied Science Private University, Amman, Jordan, for the financial support granted to this research paper.

REFERENCES

[1] S. Nashif, R. Raihan, R. Islam, and M. H. Imam, "Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System," World Journal of Engineering and Technology, vol. 6, no. 4, pp. 854-873, Nov, 2018. doi: 10.4236/wjet.2018.64057.

[2] J. S. Sonawane and D. R. Patil, "Prediction of heart disease using multilayer perceptron neural network," International Conference on Information Communication and Embedded Systems (ICICES2014), pp. 1–6, Feb, 2014. doi: 10.1109/ICICES.2014.7033860.

[3] S. A. Pattekari and A. Parveen, "Prediction system for heart disease using Naive Bayes," International Journal of Advanced Computer and Mathematical Sciences, vol. 3, no. 3, pp. 290–294, 2012.

[4] K. H. Miao, J. H. Miao, and G. J. Miao, "Diagnosing Coronary Heart Disease Using Ensemble Machine Learning," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 7, no. 10, pp. 30–39, 2016. doi: 10.14569/IJACSA.2016.071004.

[5] K. H. Miao and J. H. Miao, "Coronary Heart Disease Diagnosis using Deep Neural Networks," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 9, no. 10, pp. 1-8, 2018. doi: 10.14569/IJACSA.2018.091001.

[6] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," Expert Systems with Applications, vol. 36, no. 4, pp. 7675–7680, May, 2009. doi: https://doi.org/10.1016/j.eswa.2008.09.013.

[7] S. Bashir, U. Qamar, F. H. Khan, and M. Y. Javed, "An Efficient Rule-Based Classification of Diabetes Using ID3, C4.5, & CART Ensembles," in 2014 12th International Conference on Frontiers of Information Technology, pp. 226–231, Dec 2014. doi: 10.1109/FIT.2014.50.

[8] T. Panch, P. Szolovits, and R. Atun, "Artificial intelligence, machine learning and health systems," Journal of Global Health, vol. 8, December. 2018. doi: doi: 10.7189/jogh.08.020303.

[9] G. D. Magoulas and A. Prentza, "Machine Learning in Medical Applications," Machine Learning and Its Applications, vol. 2049, pp. 300–307, 2001. doi: https://doi.org/10.1007/3-540-44673-7_19

[10] S. A. Alasadi and W. Bhaya , "Review of data preprocessing techniques in data mining," Journal of Engineering and Applied Sciences, vol. 12, no. 16, pp. 4102–4107, Sep 2017.

[11] A. Gupta and V. Khathuria," Framingham heart study," International Journal on Future Revolution in Computer Science & Communication Engineering, vol. 4, no. 11, pp. 55–58, 2018.

[12] K.V. Nagendra and M. Ussenaiah, "Analysis of classification algorithms on heart diseases data using association rule mining, " International Journal of Computational Engineering Research(IJCER), vol. 08, no. 6, pp. 39–46, 2018.

[13] K. V. Nagendra and M. Ussenaiah, "Support vector machine and neural network classification improved by bagging," International Journal on Future Revolution in Computer Science & Communication Engineering, vol. 4, no. 2, pp. 125–130, 2018.

[14] J. Beunza, E. Puertas, E. García-Ovejero, G. Villalba, E. Condes, G. Condes, and M. F. Landecho "Comparison of machine learning algorithms for clinical event prediction(risk of coronary heart disease)," Journal of Biomedical Informatics, vol. 97, p. 103257, Sep.2017. doi:https://doi.org/10.1016/j.jbi.2019.103257.

[15] A. S. T. Nishadi, "Predicting heart diseases in logistic regression of machine learning algorithms by python jupyterlab, " International Journal of Advanced Research and Publications, vol. 3,no. 8, pp. 69–74, Aug 2019.

[16] A. Bhardwaj, A. Kundra, B. Gandhi, S. Kumar, A. Rehalia, and M. Gupta, "Prediction of heart attack using machine learning," IITM Journal of Management and IT, vol. 10, no. 1, pp. 20–24, 2019.

[17] A. Valle, A. Cinaud, V. Blachier, H. Lelong, M. E. Safar, and J. Blacher "Coronary heart disease diagnosis by artificial neural networks including aortic pulse wave velocity index and clinical parameters," Journal of Hypertension, vol. 37, no. 8, pp. 1682–1688, Aug. 2019. doi:10.1097/HJH.0000000000002075.

[18] K. Lim, B. M. Lee, U. Kang, and Y. Lee "An optimized DBN-based coronary heart disease risk prediction," International Journal of Computers Communications & Control, vol. 13,no. 4, pp. 492–502, Jul 2018. doi: https://doi.org/10.15837/ijccc.2018.4.3269 .

[19] J.K. Kim and S. Kang, "Neural network-based coronary heart disease risk prediction using feature correlation analysis, " Journal of Healthcare Engineering, vol. 2017, pp. 1–13, 2017. doi: https://doi.org/10.1155/2017/2780501.

[20] N.S. Rajliwall, G. Chetty, and R. Davey, "Chronic disease risk monitoring based on an innovative predictive modelling framework," 2017 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1–8, 2017. doi: 10.1109/SSCI.2017.8285257, 1-8, 2017.

[21] N. S. Rajliwall, R. Davey, and G. Chetty, "Machine learning based models for cardiovascular risk prediction," 2018 International Conference on Machine Learning and Data Engineering(iCMLDE), pp. 142–148, 2018. doi: 10.1109/iCMLDE.2018.00034, 142– 148, 2018.

[22] I. D. Mienye, Y. Sun, and Z Wang, "Improved sparse autoencoder based artificial neural network approach for prediction of heart disease,", Informatics in Medicine Unlocked, vol. 18, 100307, 2020.

[23] P. Puvar, N. Patel, A. Shah, R. Solanki, and D. Rana, "Heart Disease Detection using Ensemble Learning Approach," International Research Journal of Engineering and Technology (IRJET), vol. 8, no. 5, pp. 1414-1418, May, 2021.

[24] N. K. Sharma, M. Vemula, and V. Tadiboyina, "An Experimental Study of Heart Disease Prediction Using Different Supervised Machine Learning Algorithms", International Journal of Engineering Research and Technology, vol. 14, no. 3, pp. 227-240, 2021.

[25] H. A. G. Elsayed and L. Syed, "An automatic early risk classification of hard coronary heart diseases using framingham scoring model," in Proceedings of the Second International Conference on Internet of things, Data and Cloud Computing. ACM, pp. 1-8, Mar 2017. doi: https://doi.org/10.1145/3018896.3036384

[26] P. Jonsson and C. Wohlin, "An evaluation of k-nearest neighbour imputation using likertdata," 10th International Symposium on Software Metrics, 2004. Proceedings.IEEE, pp. 108-118, 2004. doi: 10.1109/METRIC.2004.1357895.

[27] G. Aksu, C. O. Güzeller, and M.T Eser, "The Effect of the Normalization Method Used in Different Sample Sizes on the Success of Artificial Neural Network Model", International Journal of Assessment Tools in Education, vol. 6, no. 2, , pp. 170–192, 2019. doi: https://doi.org/10.21449/ijate.479404.

[28] S. Prasad, "Some notes on z- scores and t- scores," International Journal of scientific research and management (IJSRM), vol. 3, no. 4, pp. 2608-2610, 2015.

[29] K. Potdar, T. S. pardawala, and C. D. pai, "A comparative study of categorical variable encoding techniques for neural network classifiers," International Journal of Computer Applications, vol. 175 ,no.4, pp. 7–9, Oct, 2017.

[30] H. LIU, F. Hussain, C. L.TAN, and M. DASH, "Discretization: An Enabling Technique," Data Mining and Knowledge Discovery, vol.6, pp. 393–423, 2002. doi: https://doi.org/10.1023/A:1016304305535.

[31] D. Ramesh, P. Suraj, and L. Saini, "Big data analytics in healthcare: A survey approach," in 2016 International Conference on Microelectronics, Computing and Communications (Mi-croCom), pp. 1–6, Jan, 2016.

[32] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Informatics in Medicine Unlocked, vol. 16, p.100203, 2019. doi: https://doi.org/10.1016/j.imu.2019.100203.

[33] A. H. Khaleel, G. A. Al-Suhail, and B. M. Hussan, "A weighted voting of k-nearest neighbor algorithm for diabetes mellitus," International Journal of Computer Science and Mobile Computing, vol. 6, no. 1, pp. 43–51, 2017.

[34] N. Ali, D. Neagu, and P. Trundle, "Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets," SN Applied Sciences, vol 1, no. 12, Nov, 2019.

[35] Y. Zhang, Z. Lin, Y. Kang, R. Ning, and Y. Meng, "A feed-forward neural network model for the accurate prediction of diabetes mellitus," INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH, vol.7, no.8, pp. 151–155, Aug, 2018.

[36] H. Yan, et al., "A multilayer perceptron-based medical decision support system for heart disease diagnosis," Expert Systems with Applications, vol.30, no.2, pp. 272 – 281, 2006.

[37] A. Olgac and B. Karlik, "Performance analysis of various activation

functions in generalized mlp architectures of neural networks," International Journal of Artificial Intelligence And Expert Systems, vol. 1, pp. 111–122, Feb, 2011.

[38] A. Smith, F. Gu, and A.D. Ball, "An Approach to Reducing Input Parameter Volume for Fault Classifiers," International Journal of Automation and Computing, vol. 16, no. 2, pp. 199-212, Apr, 2019. doi: 10.1007/s11633-018-1162-7.

[39] P. RefaeiNzadeh, L. Tang, and H. Liu, Cross-Validation.Boston, MA: Springer US, pp.532–538, 1927-2010, 2009. doi: https://doi.org/10.1007/978-0-387-39940-9_565.

[40] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," Proceedings of the 14th International Joint Conference on Artificial Intelligence, Vol. 2, pp. 1137–1143, Aug, 1995.

[41] A. Kori, "Comparative study of data classifiers using rapidminer," International Journal of Engineering Development and Research, vol. 5, pp. 1041–1043, 2017.

[42] W. Chen, S. Chen, H. Zhang, and T. Wu, "A hybrid prediction model for type 2 diabetes using k-means and decision tree," in 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), pp. 386–390, Nov 2017. doi: 10.1109/ICSESS.2017.8342938.

[43] M.Hossin and N.Sulaiman, "A review on evaluation metrics for data classification evaluations," International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 2, pp. 01–11, Mar 2015.