Machine Learning: Assisted Cardiovascular Diseases Diagnosis

Aseel Alfaidi[1], Reem Aljuhani[2], Bushra Alshehri[3], Hajer Alwadei[4], Sahar Sabbeh[5]
Department of Computer Science and Artificial Intelligence, University of Jeddah, Jeddah, KSA[1, 2, 3, 4, 5]
Faculty of Computer Sciences and Artificial Intelligence, Benha University, Egypt[5]

Abstract—**Detecting cardiovascular problems during their early stages is one of the great difficulties facing physicians. Cardiovascular diseases contribute to the deaths of around 18 million patients every year worldwide. That's why heart disease is a critical worry that must be addressed. However, it can be difficult to detect heart disease because of the multiple factors that affect health, such as high blood pressure, increased cholesterol, abnormal pulse rate, and many other factors. Therefore, the field of artificial intelligence can be instrumental in detecting diseases early on and finding an appropriate solution. This paper proposes a model for diagnosing the probability of an individual having cardiovascular illness by employing Machine Learning (ML) models. The experiments were executed using seven algorithms, and a public dataset of cardiovascular disease was used to train the models. A Chi-square test was used to identify the most important features to predict cardiovascular disease. The experiment results showed that Multi-Layer Perceptron gives the highest accuracy of disease prediction at 87.23%.**

Keywords—Cardiovascular diseases; artificial intelligence; prediction; multi-layer perceptron

I. Introduction

The human heart is the most critical component in the body, whose main task is to pump blood to all body parts. The heart is at the center of the circulatory system and is a network of blood vessels such as arteries, veins, and capillaries [1]. Any component in the body is exposed to diseases and injuries, but the heart is among the body's major organs. As its damage may threaten human life, and its diseases or injuries cannot be easily overlooked.

Heart disease disrupts the heart's regular electrical system and pumping functions. Shortness of breath, physical weakness, swollen feet, and weariness can indicate heart disease [2]. Causes threatening human heart health include high cholesterol, smoking, lack of physical activity, and increased blood pressure [3].

Cardiovascular diseases are a group of disorders brought on by cardiac issues. According to the World Health Organization [4], the leading reason for death is the cardiovascular disease as it causes the death of 18 million patients each year, around 32% of the deaths around the world. Hence, cardiovascular diseases are viewed as a significant health concern. There are several types of cardiac illness, the most prevalent are heart.

Angiography is the method that most doctors use to diagnose cardiovascular patients. However, this diagnosing process requires analyzing many factors, which is also considered an expensive procedure, especially, in developing countries that suffer from a scarcity of diagnostic devices, doctors, and other resources [5][6].

As the number of deaths caused by cardiovascular diseases rises every day, the prediction of these diseases has become one of the most crucial subjects in the medical field. Prediction helps to detect disease in its early stages, thus, reduce the risk of sickness, or treat it most effectively.

Machine learning (ML) techniques in the domain of medical diagnosis are continuously expanding. This can be attributed mostly to advancements in disease classification and recognition, which can provide data that support medical specialists in the early discovery and diagnosis of diseases, thus maintaining human health and reducing the death rate. The classification algorithms are ML learning approaches that are often used to identify the probability of disease occurrence [7] [8]. Therefore, this paper aims to build a classification model to predict cardiovascular disease using real world dataset of cardiovascular patients.

The focus of ML is to develop systems that can make predictions based on experience [8].There are three types of machine learning techniques. First, the supervised learning, where the model is trained using labeled data, and the performance of the model is evaluated using test data. Supervised learning usually includes classification and regression problems. The second type is unsupervised learning, in which the data is not labelled and the model tries to discover the hidden patterns that may exist in the data. It derives conclusions from datasets to characterize hidden knowledge after exploring data. A clustering approach is an example of unsupervised learning [9]. The third type is reinforcement learning, which neither makes use of labelled data, nor the findings are related to the data. It is concerned with how intelligent agents can take actions in an environment [10].

The classification algorithms are ML approaches that are often used to identify the probability of disease occurrence[7][8]. It is a prominent machine learning technique that uses a model inferred from training data to predict the class of new samples [11][12]. Also, classification is a supervised learning concept that categorizes a set of data into classes [12]. This paper aims to build a classification model to predict cardiovascular disease using real world dataset of cardiovascular patients.

In this paper, we applied seven different classification algorithms to predict cardiovascular disease and determine the best algorithm among them, namely, logistic regression (LR), decision tree, random forest (RF), naïve Bayesian (NB), k-nearest neighbor (KNN), support vector machine (SVM), and multi-layer perceptron (MLP).

The rest of this paper is organized as follows: Section 2 presents the related works in this filed. Section 3 describes in detail the research methodology, datasets, data preprocessing, and data analysis. Section 4 presents and discusses the results. Section 5 concludes the paper and provides the scope of future work.

II. Related Work

Researchers have suggested several possible ways to predict heart disease using various machine learning algorithms. The Cleveland Heart Disease dataset is the most common dataset used in heart disease prediction papers that are presented in this literature.

Rani et al. [13] proposed a decision system utilizing machine learning for cardio disease prediction based on a patient's clinical parameters. Their results indicated that the RF model had the best accuracy at 86.60%. Motarwar et al. [14] proposed a framework to predict the possibility of heart disease using various algorithms. They also found that RF achieved the best accuracy at 95.08%. Shah et al. [7] used several methods, such as the DT, NB, KNN, and RF algorithms. The results showed that the KNN algorithm had the greatest accuracy score at 90.7%.

Vijayashreea et al. [15] suggested a new fitness function for particle swarm optimization (PSO) using SVM. They created a new function based on identifying optimal weight population diversity and tuning for determining optimal weights. The SVM classifier's high performance was demonstrated using Receiver Operating Characteristic (ROC) analysis. In addition, they demonstrated the application of the suggested PSO-SVM–based feature selection technique for predicting heart disease. In addition, the SVM classifier was compared to other well-known classifier methods, such as NB, RF, and MLP.

Atallah and Al-Mousa [16] suggested using the complex voting ensemble method, and the outcome of the predictions is determined by a majority vote among all models. Consequently, the model attained 90% accuracy, which successfully exceeded the accuracy of each classifier. Besides the Cleveland Heart Disease dataset, Rao et al. [17] used the Switzerland, Hungarian, and Long Beach datasets. Using different algorithms for each dataset, the results showed that the highest accuracies were 86.81%, 98.30%, 84.26%, and 82.20% for Cleveland, Switzerland, Hungarian, and Long Beach, respectively.

Mohan et al. [18] developed a method that aims to improve the accuracy of cardiovascular disease prediction by applying machine learning techniques to find essential features. Feature selection depended on the machine learning techniques used, which included NB, generalized linear models, linear regression, deep learning, DT, RF, and SVM. The proposed hybrid random forest and linear model method was shown to be very accurate at predicting heart disease, with an accuracy of 88.7%. Another hybrid predictive system was proposed by Haq et al. [19] to diagnose heart disease. The authors used seven popular machine learning algorithms: LR, KNN, ANN, SVM, NB, DT, and RF. As a result, they concluded that LR had the best accuracy for predicting heart disease, with an accuracy of 89%. Kavitha et al. [20] proposed a heart disease prediction model using a hybrid approach. They implemented the model using three machine learning algorithms: RF, DT, and a hybrid of the two. Results showed a highest accuracy of 88.7%, achieved by the hybrid model.

Lakshmanarao et al. [21] suggested heart disease prediction using an ensemble classifier model. Two datasets were used in their study. First, they applied two feature selection techniques, namely Analysis of Variance (ANOVA) for F-value and mutual information. Based on these two techniques, they determined the best features. Nowshad et al. [22] gathered data in Bangladesh's Sylhet district by visiting hospitals and healthcare businesses in person to create a good questionnaire for heart disease prediction. There are 564 instances and 18 attributes in their dataset. The SVM produced the best results, with an accuracy level of 91%.

In contrast to previous studies, we used the cardiovascular disease dataset. To the best of our knowledge, this is the first study using this dataset which includes 70000 patients and 11 features. Also, we have applied different ML algorithms to determine the best for obtaining precise results for predicting cardiovascular disease.

III. Research Methodology

This section presents the methodology followed by researchers for the prediction of cardiovascular disease using machine learning algorithms. Fig. 1 illustrates our methodology process flow. First data are pre-processed, most informative features are selected, the resulting data are fed into different classification models and finally, performance is evaluated.

A. Dataset

The Cardiovascular Disease dataset obtained from the Kaggle repository [23] was used. The dataset has a sample size of 70000 patients and 11 features. Table I displays the feature dataset details and an explanation of each feature.

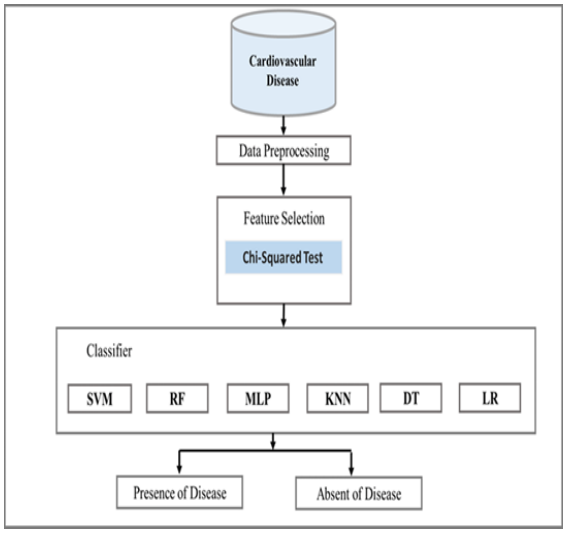

Fig. 1. Methodology Proposed to Predict Cardiovascular Disease.

TABLE I. INFORMATION IN THE DATASET

Feature name	Type	Description
Age	Discrete	Number of days
Gender	Discrete	Female: 2, Male: 1
Height	Continuous	In cm, Max = 250, Min = 55
Weight	Continuous	In kg, Max = 200, Min = 10
Systolic blood pressure	Discrete	Max = 16020, Min = −150
Diastolic blood pressure	Discrete	Max =11000, Min = −70
Cholesterol	Discrete	1: normal, 2: above normal, 3: well above normal
Glucose	Discrete	1: normal, 2: above normal, 3: well above normal
Smoking	Discrete	present: 1, absent: 0
Alcohol intake	Discrete	present: 1, absent: 0
Physical activity	Discrete	present: 1, absent: 0
Cardiovascular disease	Discrete	present: 1, absent: 0

B. Data Pre-processing

The Pre-processing of the dataset for a machine learning model is necessary for efficiency. We suggest in this study the pre-processing techniques of removing anomalies (outliers) and applying standard scaler to the dataset for showing the models efficiency and obtaining an acceptable and reliable accuracy for predicting the disease. After that, we used 68733 records from the dataset. We also modified some features of the dataset to best identify the factors that most influence cardiovascular disease as follows:

- Weight and height were merged into one feature in Body Mass Index (BMI): calculates body fat percentage based on height and weight.
- Features of the dataset were transformed while maintaining the information to make it more comprehensible. In this dataset, age was converted from days to years, and the gender feature was converted to binary.
- Values out of range (outliers) were removed in the.

C. Explanatory Data Analytics

The explanatory data analytics aims to use statistical and/or visual techniques to get insights into data sparsity, correlation, distribution, …etc.

The pie chart in Fig. 2A displays the gender distribution of the dataset, male 65.13% and female 34.87%. Fig. 2B shows the relationship of the gender feature to disease, indicating that the average number of females with cardiovascular disease is more than that of males.

Fig. 3 shows features relationships with the target feature. Fig. 3A indicates the cholesterol feature. Cholesterol is a waxy substance found in the blood, and high blood cholesterol is one of the heart diseases factors [24]. In the dataset, the rates of infection range between 1) normal and 3) well above normal; 30 well above normal has the highest occurrence of disease.

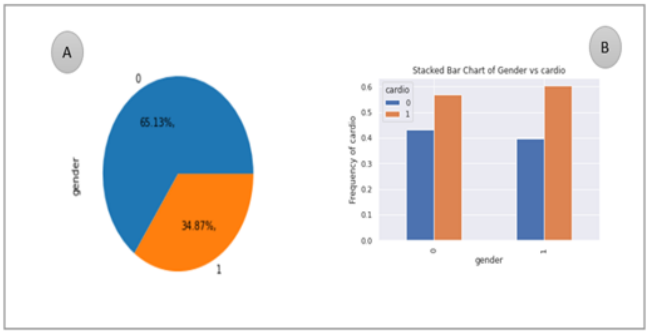

Fig. 2. Gender Feature Distribution of the Dataset.

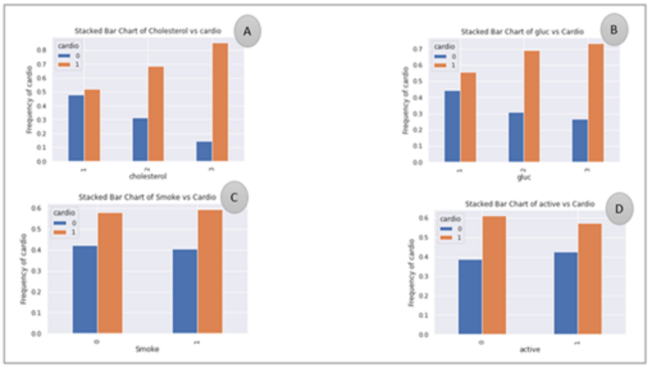

Fig. 3. Feature Distribution.

Fig. 3B shows the relationship between glucose and cardiovascular disease. High glucose is a known factor in cardiovascular disease [24]. The rates of glucose range between 1: normal and 3: well above normal in the dataset. Cardiovascular disease is most highly represented in 3: well above normal.

Indicated in Fig. 3C is the effect of smoking on the heart, a strongly linked known factor that causes cardiovascular disease [24]. The dataset indicates that it may have an effect in some cases. Finally, in Fig. 3D, it is shown that physical activity has a little effect on cardiovascular disease.

As demonstrated in Fig. 4, data visualization is utilized for discrete features to preview the distribution in the data. As shown in Fig. 4A, age data is distributed between 35 and 65 years. For the systolic blood pressure (ap_hi) attribute in Fig. 3B, we note a distribution from less than 100 to 200. Systolic blood pressure represents the heart's force on artery walls each time it beats [25]. We present the diastolic blood pressure (ap_lo) attribute in Fig. 4C, distributed between 50 and 125. Diastolic blood pressure measures the pressure on the walls of arteries between heartbeats [26].

Between each feature and the target feature(cardio), a correlation value was determined as shown in Table II. We note that the features that are positively correlated with the target feature (cardio) are ap_hi, ap_lo, age, cholesterol, BMI, gluc, gender, alco, and smoke; the active feature is the negatively correlated feature with the target feature (cardio).

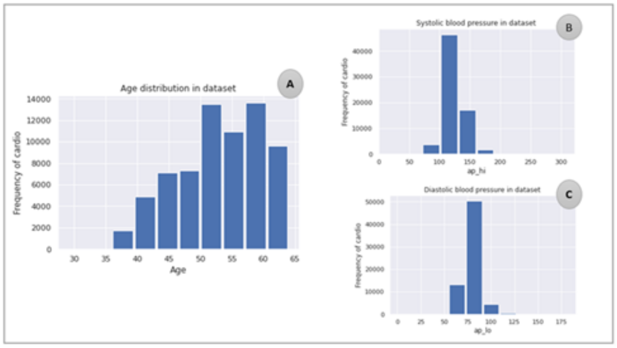

Fig. 4. Discrete Feature Distribution.

TABLE II. FEATURE AND CORRELATION VALUE OF THE DATASET

Feature code	Correlation value
Ap_hi	0.625865
Ap_lo	0.541238
Age	0.268197
Cholesterol	0.229901
BMI	0.217294
Gluc	0.111875
Gender	0.033391
Alco	0.018351
Smoke	0.008398
Id	0.004876
Active	-0.031358

D. *Feature Selection*

Feature selection is an essential step before data is fed to the classification model. It aims to reduce dimensionality by selecting the most informative feature that can contribute positively to the performance of the models.

In this study, for feature selection, we applied the filter method that uses variable ranking approaches as the primary criterion for ordering variables [27]. It is also a popular feature selection method in machine learning techniques and has achieved success for practical applications [27]. Moreover, we selected from the Chi-Squared test method. It uses statistical techniques to evaluate the relationship between the features and the target variables. In addition, it is used when a feature is tested and the target variable in the classification problem [27].

E. *Machine Learning Models*

For our experiment, we applied seven machine learning techniques namely, logistic regression, decision tree, random forest, naïve Bayesian, k-nearest neighbors, support vector machine, and multi-layer perceptron.

1) Logistic Regression (LR): a predictive analysis technique based on the concept of probability. LR can be compared to the Linear Regression model, LR, on the other hand, employs a more complicated computation: the sigmoid function, commonly known as a logistic function. In the context of LR, the cost function's range is from zero to one. Linear functions cannot be employed since their values can be less than zero or larger than one [22].

2) Random Forest (RF): The RF algorithm is a supervised classification approach. In RF, a forest is formed by several trees, each of which emits a class expectation, with the class with the most votes becoming the model's forecast. The bigger the number of trees in the RF classifier, the more accurate it is. It can be used for a variety of tasks, including classification and regression, but it shines when it comes to classification and dealing with missing information [28].

3) Decision Tree (DT): This is a classification algorithm that can be used to classify both category and numerical data. It has a type of structure that resembles a tree. DT is a primary and commonly used method for dealing with medical data. The data in a tree-shaped graph is simple to build and analyze. We used the DT classification method since it is one of the best and most widely used controlled learning strategies[29]. It is simple to build a stable decision tree for a given data collection.

4) Naïve Bayes (NB): We also use the NB classifier, a machine learning approach for identifying and forecasting the probability of an occurrence, is also used. Every NB classifier presupposes that a features value is different from the values of any other features in the class variables [22].

5) K-Nearest Neighbour (KNN): That is one of the most fundamental regression and classification machine learning approaches. The data is employed in KNN calculations, which use resemblance measures to characterize new points (e.g., distance function). In a nutshell, the KNN computation assumes the closeness of the comparable objects. In KNN, classification is made by taking into account the majority vote of its neighbours. The data point is labelled with the class that has most of its neighbours [30]. The selection of k and the accuracy may increase as the number of nearest neighbours increases.

6) Support Vector Machine (SVM): That can be used to perform regression and classification tasks. To implement SVM, we first represent each data item in our model as a number of features, with each component's estimation corresponding to a specific coordinate. The data is then classified by determining hyper-plane, which best separates the classes [7].

7) Multi-Layer Perceptron (MLP): Recently, it has been demonstrated that neural networks, specifically MLP, are excellent alternatives to more traditional statistical methodologies. It has been demonstrated that MLP may be trained to resemble almost any smooth, measurable function [31]. Unlike other statistical techniques, MLP makes no assumptions about the distribution of data. When given new, unknown inputs, it can represent extremely nonlinear functions and be trained to generalize appropriately. These characteristics make the MLP an appealing option for constructing numerical models as well as choosing among statistical approaches [31].

F. *Evaluation Metrics*

We suggest multiple ways to evaluate the classifiers as shown in Table III, to determine the appropriate model to predict disease.

TABLE III. EVALUATION METRICS

Definition	**Equation**
Accuracy: It is an assessment of a system ability to make correct predictions.	Accuracy = $(\frac{\text{Correct predictions}}{\text{Total predictions}}) \times 100$
Sensitivity: It is an assessment measures the ability of a system to predict positive outcomes correctly.	Sensitivity = $(\frac{\text{True positives}}{\text{True positives + false negatives}})$ × 100.
Specificity : It is an assessment measures the ability of a system to predict negative outcomes correctly.	Specificity = $(\frac{\text{True negatives}}{\text{True negatives + false positives}})$ × 100.
Precision: It is an assessment measures of a system to to the relevant results.	Precision = $(\frac{\text{True positives}}{\text{True positives + false positives}})$ × 100.
F-measure: Is the sum of the results of measuring accuracy and sensitivity	F-measure = $2 \times (\frac{\text{Sensitivity} \times \text{precision}}{\text{Sensitivity + precision}})$

IV. RESULTS AND DISCUSSION

For our experiments, data were divided into training and testing sets with proportions of 70% and 30%, respectively. The classification was performed using the medical biomarkers available in the dataset, and class 1 means that the individual has a disease, while class 0 means that the person is disease-free.

Our first experiment targets the evaluating of the optimal number of features that can achieve the best accuracy among all models. To find that number, we assessed the accuracy obtained with each subset from one to ten combinations of features. Applying the Chi-squared method, the selection of features will be based on their rank that is determined by their scores. Tables IV and V shows all the combination of features and their corresponding accuracy by each classifier. As a result, with five selected features the highest accuracy was obtained by both MLP, LR, SVM, and RF with 87.2%, 85.5% 86.6%, and 86.0% respectively, whereas with all selected the highest accuracy was obtained by MLP with 87.2%.

TABLE IV. PERFORMANCE OF CLASSIFIERS ON DIFFERENT NUMBERS OF FEATURES

Number of features selected	**Accuracy obtained by each model (%)**						
	MLP	**LR**	**SVM**	**RF**	**KNN**	**NB**	**DT**
5	87.2	85.5	86.6	86	84.6	83.4	85.9
7	87.1	85.4	86.5	85.6	84.7	83.4	85.4
9	87.1	85.4	86.5	85.5	84.7	83.4	85.4

TABLE V. THE NUMBER OF FEATURES WITH CORRESPONDING SCORE

Number of features	**Name of feature**	**Score**
1	Ap_hi	59514.160186
2	Ap_hi	59514.160186
	Ap_lo	23057.693730
3	Ap_hi	59514.160186
	Ap_lo	23057.693730
	Bmi	4436.047195
4	Ap_hi	59514.160186
	Ap_lo	23057.693730
	Bmi	4436.047195
	Age	4289.087958
5	Ap_hi	59514.160186
	Ap_lo	23057.693730
	Bmi	4436.047195
	Age	4289.087958
	Cholesterol	1226.927745
6	Ap_hi	59514.160186
	Ap_lo	23057.693730
	Bmi	4436.047195
	Age	4289.087958
	Cholesterol	1226.927745
	Gluc	229.196698
7	Ap_hi	59514.160186
	Ap_lo	23057.693730
	Bmi	4436.047195
	Age	4289.087958
	Cholesterol	1226.927745
	Gluc	229.196698
	Gender	49.914749
8	Ap_hi	59514.160186
	Ap_lo	23057.693730
	Bmi	4436.047195
	Age	4289.087958
	Cholesterol	1226.927745
	Gender	49.914749
	Alco	21.911426
9	Ap_hi	59514.160186
	Ap_lo	23057.693730
	Bmi	4436.047195
	Age	4289.087958
	Cholesterol	1226.927745
	Gluc	229.196698
	Gender	49.914749
	Alco	21.911426
	Active	13.288970
10	Ap_hi	59514.160186
	Ap_lo	23057.693730
	Bmi	4436.047195
	Age	4289.087958
	Cholesterol	1226.927745
	Gluc	229.196698
	Gender	49.914749
	Alco	21.911426
	Active	13.288970
	Smoke	4.430920

The performances of classifiers were evaluated with all features successively as shown in Table IV. With all selected sets of features, the MLP classifier outperformed all the other models achieving the highest accuracy. Additionally, results showed that the highest accuracy was achieved using the top five features and choosing from six to ten features achieved slightly lower accuracy with only a 0.1% difference. Also, choosing from one to four features achieved lower accuracy with only a 1.0%. The SVM comes after with 86.6% accuracy. DT, RF, and LR almost achieve the same accuracy and NB achieved the lowest accuracy among all models.

Afterwards, the top five features were used as input to the classifiers and performance was evaluated in terms of accuracy. The first test was the LR classifier, which produced an accuracy of 85.5%. We performed the second test of the dataset for the RF classifier and a third test for the DT classifier. These classifiers achieved accuracies of 86% and 85.9%, respectively, and are approximately equal to the first classifier accuracy.

In addition, we used NB and KNN classifiers and achieved accuracies of 83.4% and 84.7%, respectively, which are less than the previous classifiers. Furthermore, we performed tests for the SVM and MLP classifiers; they achieved the highest accuracies of 86.6% and 87.2%, respectively.

We also measured the performance of classifiers for several measurements, to illustrate their robustness and prediction capability. Table VI shows a comparison of classifiers based on several measurements. In terms of the sensitivity measure, the LR classifier had 84.34%. In specificity and precision measures, the SVM classifier reached the highest rates of 95.51% and 96.06%, respectively. Finally, for the F-measure, the MLP classifier had 88.13% and the best accuracy was also achieved by MLP at 87.23%.

To further investigate the models that were selected for cardiovascular disease prediction, we display a ROC curve as shown in Fig. 5. The ROC curve is a metric for each classifier's ability. The model performs best for predicting when the area value is closer to one. It is clear to note that the ROC curve and accuracy result of MLP is the best among the other classifiers for predicting cardiovascular disease.

TABLE VI. EVALUATION PARAMETERS FOR ALL CLASSIFIERS (VALUES LISTED IN PERCENTAGES)

Classifier	Accuracy	Sensitivity	Specificity	Precision	F-measure
LR	85.54	84.34	87.18	90.01	87.09
RF	86.03	82.19	91.28	92.82	87.19
DT	85.93	81.24	92.34	93.57	86.97
NB	83.38	76.44	92.89	93.65	84.18
KNN	84.56	83.66	85.79	88.97	86.24
SVM	86.63	80.16	95.51	96.06	87.39
MLP	87.23	82.01	94.37	95.23	88.13

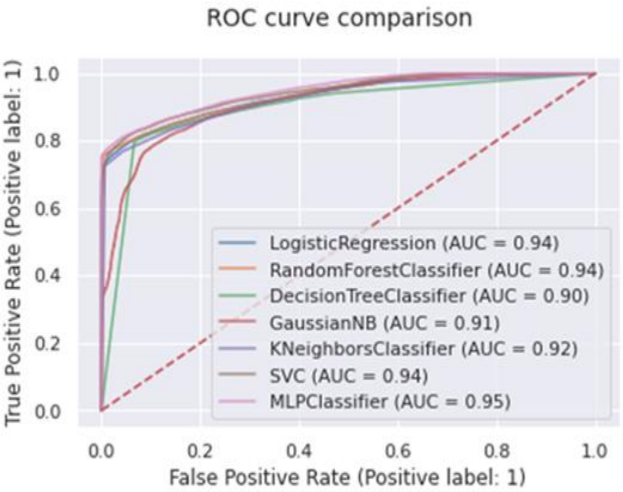

Fig. 5. ROC Curve for all Classifiers.

V. CONCLUSION AND FUTURE WORK

This research studies the performance of machine learning techniques to predict the probability of cardiovascular disease. A dataset of cardiovascular disease for 70000 patients was used for our experiment. Models' performance was evaluated in terms of their accuracies. we have introduced some steps for pre-processing the dataset. In addition, we chose more informative features that impact the performance of the models. Results show that the MLP model demonstrated the highest accuracy in predicting cardiovascular disease.

For our future work, different feature selection techniques can be used to explore the best. More datasets can be used for better and more accurate evaluation. Finally, deep learning techniques can be applied to the prediction problem.

REFERENCES

[1] N. Heart, Lung, B. Institute, N. American, and A. for the S. of Obesity, “The practical guide: identification, evaluation, and treatment of overweight and obesity in adults,” Phys. Lett. Sect. A Gen. At. Solid State Phys., vol. 379, no. 10–11, pp. 870–872, 2000, doi: 10.1016/j.physleta.2015.01.006.

[2] D. Deng, P. Jiao, X. Ye, and L. Xia, “An image-based model of the whole human heart with detailed anatomical structure and fiber orientation,” Comput. Math. Methods Med., vol. 2012, 2012, doi: 10.1155/2012/891070.

[3] M. Elhneiti and M. Al-Hussami, “Predicting Risk Factors of Heart Disease among Jordanian Patients,” Health (Irvine. Calif)., vol. 09, no. 02, pp. 237–251, 2017, doi: 10.4236/health.2017.92016.

[4] “Cardiovascular diseases.” https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (accessed Nov. 29, 2021).

[5] R. Katarya and P. Srinivas, “Predicting Heart Disease at Early Stages using Machine Learning: A Survey,” Proc. Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2020, no. Icesc, pp. 302–305, 2020, doi: 10.1109/ICESC48915.2020.9155586.

[6] Y. Zhao, E. P. Wood, N. Mirin, S. H. Cook, and R. Chunara, “Social Determinants in Machine Learning Cardiovascular Disease Prediction Models: A Systematic Review,” Am. J. Prev. Med., vol. 61, no. 4, pp. 596–605, 2021, doi: 10.1016/j.amepre.2021.04.016.

[7] M. Diwakar, A. Tripathi, K. Joshi, M. Memoria, P. Singh, and N. Kumar, “Latest trends on heart disease prediction using machine learning and image fusion,” Mater. Today Proc., vol. 37, no. Part 2, pp. 3213–3218, 2020, doi: 10.1016/j.matpr.2020.09.078.

[8] C. J. Harrison and C. J. Sidey-Gibbons, “Machine learning in medicine: a practical introduction to natural language processing,” BMC Med. Res. Methodol., vol. 21, no. 1, pp. 1–18, 2021, doi: 10.1186/s12874-021-01347-1.

[9] M. Batta, “Machine Learning Algorithms - A Review ,” Int. J. Sci. Res. (IJ, vol. 9, no. 1, pp. 381-undefined, 2020, doi: 10.21275/ART20203995.

[10] E. F. Morales and J. H. Zaragoza, “An introduction to reinforcement learning,” Decis. Theory Model. Appl. Artif. Intell. Concepts Solut., pp. 63–80, 2011, doi: 10.4018/978-1-60960-165-2.ch004.

[11] M. Pérez-Ortiz, S. Jiménez-Fernández, P. A. Gutiérrez, E. Alexandre, C. Hervás-Martínez, and S. Salcedo-Sanz, “A review of classification problems and algorithms in renewable energy applications,” Energies, vol. 9, no. 8, pp. 1–27, 2016, doi: 10.3390/en9080607.

[12] S. Pandey, M. Supriya, and A. Shrivastava, “Data Classification Using Machine Learning Approach,” no. June, 2018, doi: 10.1007/978-3-319-68385-0.

[13] P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, “A decision support system for heart disease prediction based upon machine learning,” J. Reliab. Intell. Environ., vol. 7, no. 3, pp. 263–275, 2021, doi: 10.1007/s40860-021-00133-6.

[14] P. Motarwar, A. Duraphe, G. Suganya, and M. Premalatha, “Cognitive Approach for Heart Disease Prediction using Machine Learning,” Int. Conf. Emerg. Trends Inf. Technol. Eng. ic-ETITE 2020, 2020, doi: 10.1109/ic-ETITE47903.2020.242.

[15] J. Vijayashree and H. P. Sultana, “A Machine Learning Framework for Feature Selection in Heart Disease Classification Using Improved Particle Swarm Optimization with Support Vector Machine Classifier,” Program. Comput. Softw., vol. 44, no. 6, pp. 388–397, 2018, doi: 10.1134/S0361768818060129.

[16] R. Atallah and A. Al-Mousa, “Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method,” 2019 2nd Int. Conf. New Trends Comput. Sci. ICTCS 2019 - Proc., pp. 0–5, 2019, doi: 10.1109/ICTCS.2019.8923053.

[17] V. W. Xqdo, “Computational Analysis of Machine Learning Algorithm to predict Heart Disease,” vol. 5, pp. 960–964, 2021.

[18] S. Mohan, C. Thirumalai, and G. Srivastava, “Effective heart disease prediction using hybrid machine learning techniques,” IEEE Access, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[19] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, R. Sun, and I. Garciá-Magarinõ, “A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms,” Mob. Inf. Syst., vol. 2018, 2018, doi: 10.1155/2018/3860146.

[20] G. Renugadevi, G. Asha Priya, B. Dhivyaa Sankari, and R. Gowthamani, “Predicting heart disease using hybrid machine learning model,” J. Phys. Conf. Ser., vol. 1916, no. 1, 2021, doi: 10.1088/1742-6596/1916/1/012208.

[21] A. Lakshmanarao, A. Srisaila, and T. S. R. Kiran, “Heart disease prediction using feature selection and ensemble learning techniques,” Proc. 3rd Int. Conf. Intell. Commun. Technol. Virtual Mob. Networks, ICICV 2021, no. Icicv, pp. 994–998, 2021, doi: 10.1109/ICICV50876.2021.9388482.

[22] M. N. R. Chowdhury, E. Ahmed, M. A. D. Siddik, and A. U. Zaman, “Heart Disease Prognosis Using Machine Learning Classification Techniques,” 2021 6th Int. Conf. Converg. Technol. I2CT 2021, pp. 1–6, 2021, doi: 10.1109/I2CT51068.2021.9418181.

[23] “Cardiovascular Disease dataset | Kaggle.” https://www.kaggle.com/sulianova/cardiovascular-disease-dataset (accessed Nov. 29, 2021).

[24] K. J. Bowen, V. K. Sullivan, P. M. Kris-Etherton, and K. S. Petersen, “Nutrition and Cardiovascular Disease—an Update,” Curr. Atheroscler. Rep., vol. 20, no. 2, 2018, doi: 10.1007/s11883-018-0704-3.

[25] U.S. Department of Health and Human Services, How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease. 2010.

[26] J. Tan, X. Zhang, W. Wang, P. Yin, X. Guo, and M. Zhou, “Smoking, blood pressure, and cardiovascular disease mortality in a large cohort of chinese men with 15 years follow-up,” Int. J. Environ. Res. Public Health, vol. 15, no. 5, 2018, doi: 10.3390/ijerph15051026.

[27] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” Comput. Electr. Eng., vol. 40, no. 1, pp. 16–28, 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[28] S. Xu, Z. Zhang, D. Wang, J. Hu, X. Duan, and T. Zhu, “Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework,” 2017 IEEE 2nd Int. Conf. Big Data Anal. ICBDA 2017, pp. 228–232, 2017, doi: 10.1109/ICBDA.2017.8078813.

[29] V. Sharma, S. Yadav, and M. Gupta, “Heart Disease Prediction using Machine Learning Techniques,” Proc. - IEEE 2020 2nd Int. Conf. Adv. Comput. Commun. Control Networking, ICACCCN 2020, vol. 1, no. 6, pp. 177–181, 2020, doi: 10.1109/ICACCCN51052.2020.9362842.

[30] S. B. Imandoust and M. Bolandraftar, “Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events : Theoretical Background,” Int. J. Eng. Res. Appl., vol. 3, no. 5, pp. 605–610, 2013.

[31] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” Neural Networks, vol. 2, no. 5, pp. 359–366, 1989, doi: 10.1016/0893-6080(89)90020-8.