Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 8, S. 23-31, 6, 2024
© Telif hakkı IJANSER'e aittir
Araştırma Makalesi

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 8, pp. 23-31, 6, 2024
Copyright © 2024 IJANSER
Research Article

https://as-proceeding.com/index.php/ijanser
ISSN: 2980-0811

Comparative Analysis of Machine Learning Classification Models in Predicting Cardiovascular Disease

Ladislav Végh [1]*, Ondrej Takáč [1], Krisztina Czakóová [1], Daniel Dancsa [2] and Melinda Nagy [2]

[1] *Department of Informatics, Faculty of Economics and Informatics, J. Selye University, Slovakia*
[2] *Department of Biology, Faculty of Education, J. Selye University, Slovakia*

Email of the corresponding author: (veghl@ujs.sk)

(Received: 06 July 2024, Accepted: 24 July 2024)

(4th International Conference on Scientific and Academic Research ICSAR 2024, July 19 - 20, 2024)

ATIF/REFERENCE: Végh, L., Takáč, O., Czakóová, K., Dancsa, D. & Nagy, M. (2024). Comparative Analysis of Machine Learning Classification Models in Predicting Cardiovascular Disease. *International Journal of Advanced Natural Sciences and Engineering Researches*, 8(6), 23-31.

Abstract – For a long time, cardiovascular diseases have been the leading cause of death worldwide. Machine learning has found significant usage in the medical field as it can find patterns in data. Classification models can help cardiologists to diagnose heart diseases and minimize misdiagnosis accurately. In this paper, we explored a dataset related to heart disease and compared the accuracy of 43 machine learning classification models. The dataset for this research was downloaded from Kaggle; it contained 1190 observations, 11 features (age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiogram results, maximum heart rate achieved, exercise induced angina, oldpeak, the slope of the peak exercise ST segment) and a binary target variable (no heart disease or observed cardiovascular disease). For data exploration, preprocessing, training, testing, and predictor importance analysis, we used MATLAB R2004a software and the Classification Learner app included in this software. Before training machine learning classification models, we divided the dataset into a training set (90% of observations) and a test set (10% of observations). To prevent overfitting during the training of classification models, 10-fold cross-validation was used. The result showed that the best accuracy was reached with an optimized ensemble classification model (validation accuracy: 0.9262 and test accuracy: 0.9580). After calculating the permutation importance of each feature, we observed that the most important feature among all 11 features was the slope of the peak exercise ST segment.

Keywords – Data Analysis, Machine Learning, Classification Models, Heart Disease, Cardiovascular Disease.

I. INTRODUCTION

Cardiovascular illnesses have been the world's top cause of death for many years; 20.5 million deaths worldwide in 2021 were related to cardiovascular disease [1]. Cardiovascular disease is a medical condition that impacts the heart and blood arteries, potentially affecting several body regions. The issues encompass vasoconstriction, congenital cardiac and vascular abnormalities, valvular dysfunction, and arrhythmias [2]. Accurate diagnosis of cardiovascular illnesses is essential for cardiologists to deliver appropriate treatment.

Using machine learning in various fields has grown due to its ability to recognize patterns from data [3]–[5]. Machine learning can be successfully utilized in the medical field, as well. Using machine learning to classify cardiovascular disease occurrence can aid diagnosticians in minimizing misdiagnosis [6][7]. In this paper, we first analyzed a dataset related to heart disease and then compared the accuracy, precision, recall, and f1 score of several machine learning classification models.

II. MATERIALS AND METHOD

We used dataset [8][9] downloaded from Kaggle. The MATLAB R2024a [10] software and its Classification Learner app [11] were used for data preprocessing, data analysis, and comparing several classification models.

A. *Dataset*

The dataset [8][9] contained 1190 observations and 12 features, including the target variable (age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiogram results, maximum heart rate achieved, exercise induced angina, oldpeak, the slope of the peak exercise ST segment, target). The binary target variable had the following values: 1 for observed cardiovascular disease (heart disease) and 0 for no heart disease (normal).

B. *Data Preprocessing*

During data examination, we noticed several observations contained some incorrect values. There were 1 zero value for resting blood pressure, 172 zero values for serum cholesterol, and 13 negative values for oldpeak. We replaced these incorrect values with the corresponding mean values: 132.264929 for the resting blood pressure, 245.906680 for the serum cholesterol, and 0.943840 for the oldpeak.

C. *Dataset Analysis*

First, we examined the distribution of the target variable. In the left chart of Fig. 1, we can see that 53% of all observations had a value of 1 (heart disease), and 47% had a value of 0 (normal). This distribution of values of the target variable is suitable for training and testing machine learning classification models without further data processing, as the dataset is not imbalanced.

Next, in the following part of this subchapter, we examined 11 dataset features. For every feature, we compared the observations with heart disease (red on the following graphs) to those without cardiovascular disease (green color on the following charts).

The right chart of Fig. 1 shows the distribution by age. Hearth disease was observed more often in older people.

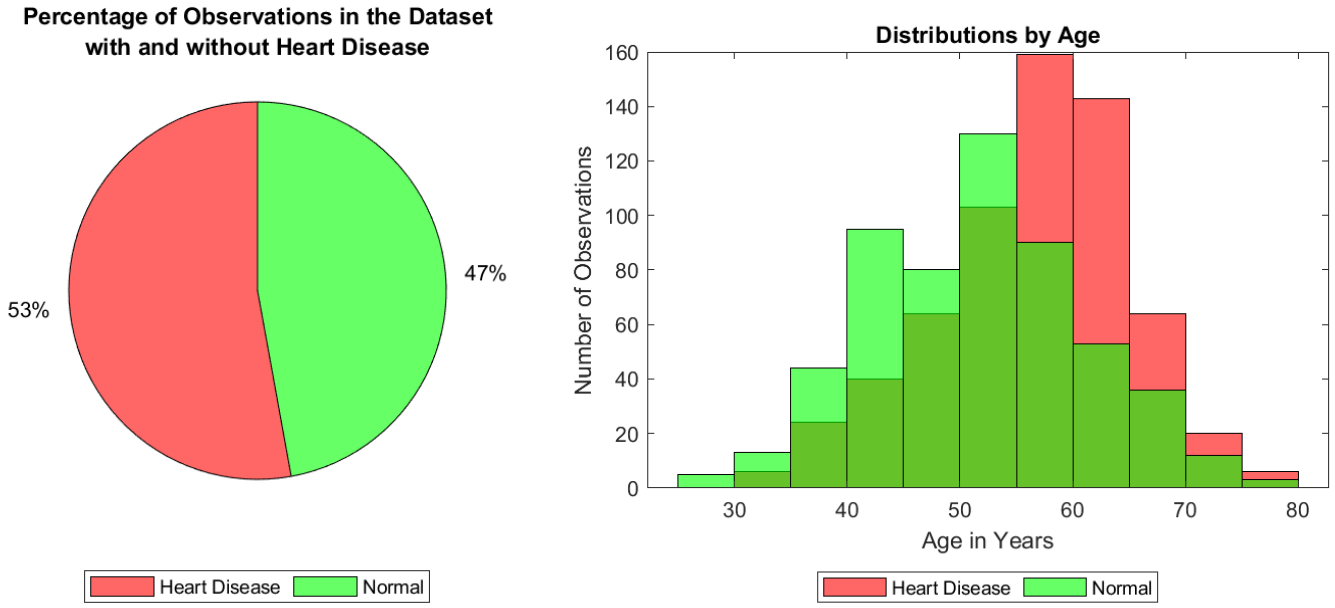

Fig. 1 Distribution by target variable (left) and age (right)

Fig. 2 illustrates the distribution of the target variable by gender (left) and chest paint type (right). In the left charts, we can see that heart disease was diagnosed more often among males than females. The right chart clearly shows that asymptomatic chest pain type was detected more often among observations with heart disease.

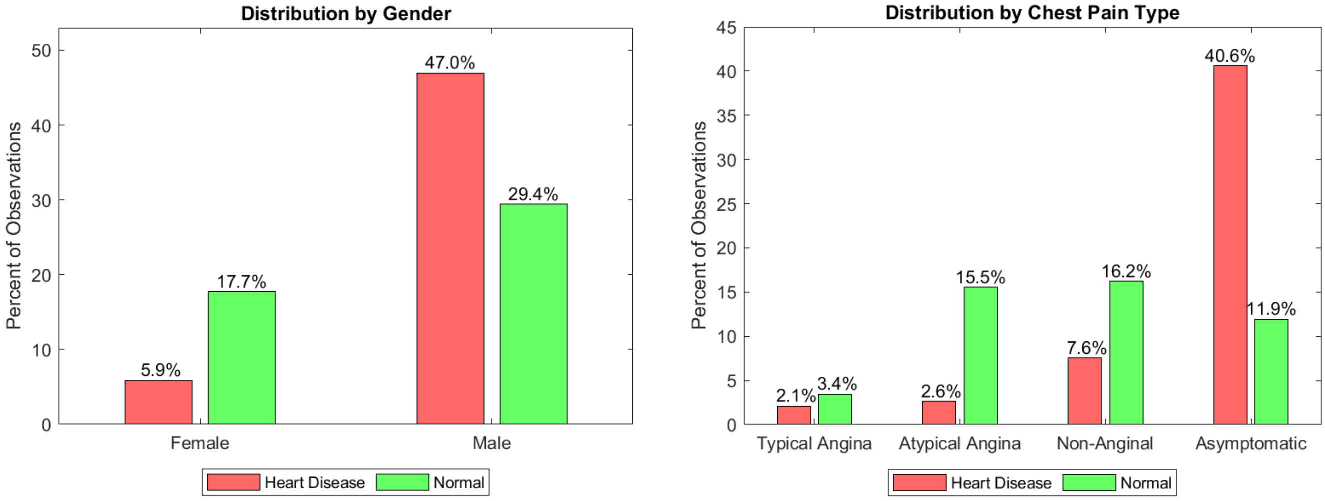

Fig. 2 Distribution by gender (left) and chest pain type (right)

The charts in Fig. 3 illustrate the data distribution by resting blood pressure and serum cholesterol. These charts do not show many differences between observations with heart disease and observations without heart disease in this dataset.

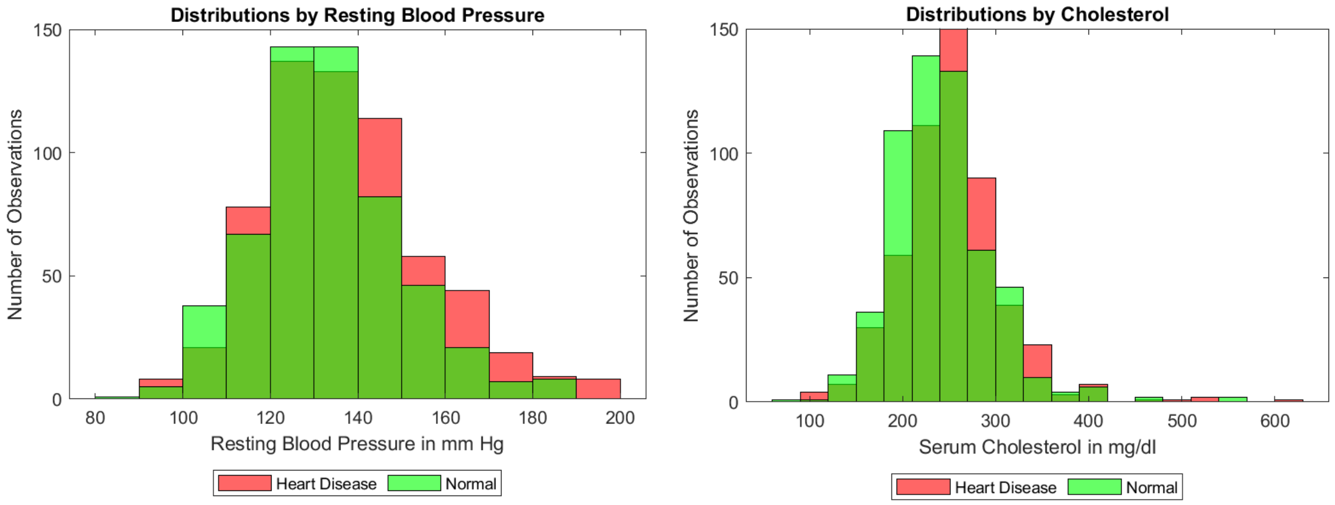

Fig. 3 Distribution by resting blood pressure (left) and serum cholesterol (right)

The left chart in Fig. 4 shows the distribution of the target variable by fasting blood sugar. This chart shows that heart disease appeared more often in the dataset among observations where the fasting blood sugar value was more than 120 mg/dl than among observations with less fasting blood sugar value.

The distribution by resting electrocardiogram (ECG) results is shown in the right chart of Fig. 4. The resting ECG result had three values in the dataset: value 0 for normal resting ECG, value 1 for resting ECG having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), and value 2 for resting ECG showing probable or definite left ventricular hypertrophy by Estes' criteria [9]. In the right chart in Fig. 4, we can observe that cardiovascular disease more often occurred when resting ECG had ST-T wave abnormality or resting ECG showed left ventricular hypertrophy, and heart disease less often occurred with normal resting ECG.

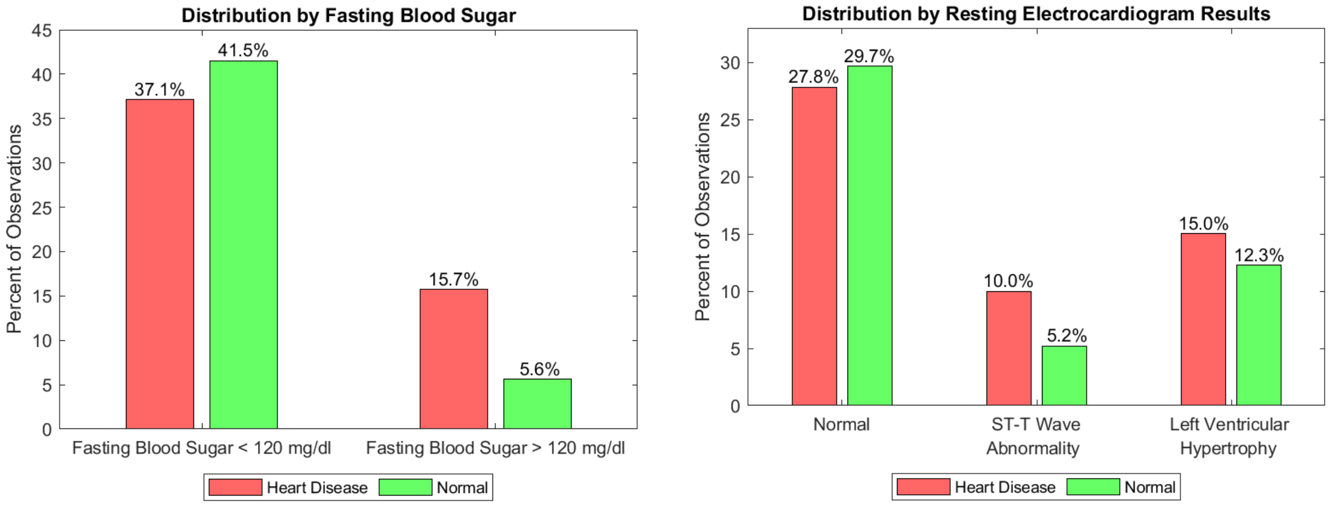

Fig. 4 Distribution by fasting blood sugar (left) and resting electrocardiogram results (right)

The following graph (left chart in Fig. 5) shows the distribution of observations by maximum heart rate. The chart indicates that observations with heart disease (red in the histogram) usually have lower maximum heart rates than observations without cardiovascular disease (green color in the histogram).

The right chart of Fig. 5 illustrates the distribution of the target variable by exercise induced angina. The graph shows that among observations where the angina was induced by exercise, heart disease appeared significantly more often.

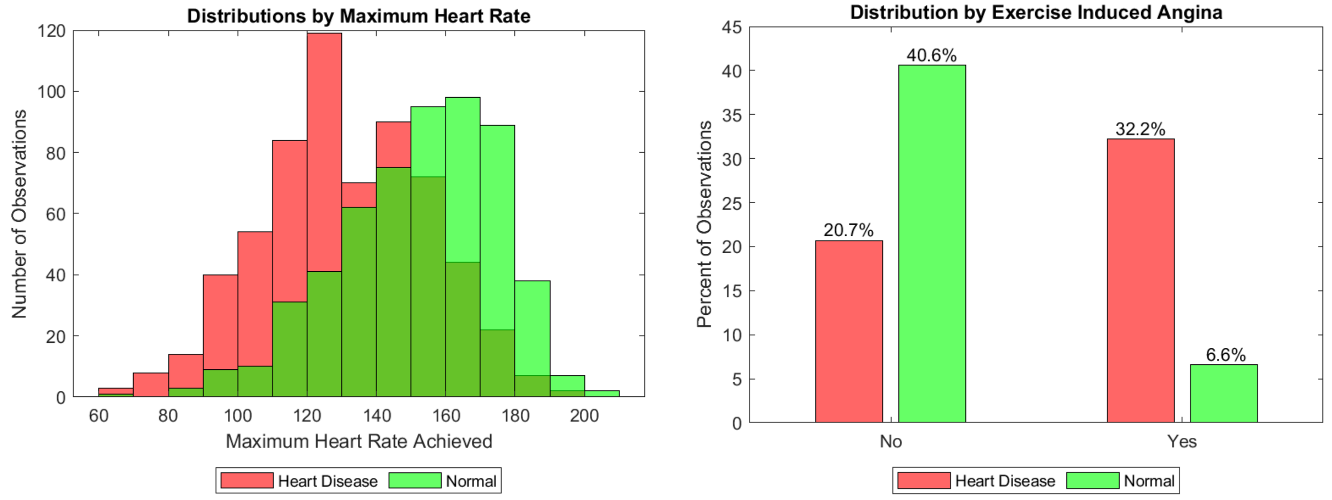

Fig. 5 Distribution by maximum heart rate (left) and exercise induced angina (right)

The following two graphs in Fig. 6 show the distribution of observations by oldpeak (left chart) and the distribution of the target variable by the slope of the peak exercise ST segment (right chart). Both graphs show differences between observations with and without heart disease. In the left chart, we can see that the oldpeak usually has a higher value in observations with heart disease (red color) than in cases where no cardiovascular disease was observed (green color). The right chart shows that most of the observations with heart disease (red color) have flat slopes of the peak exercise ST segment, while most of the cases where no cardiovascular disease was observed (green color) have upsloping slopes of the peak exercise ST segment.

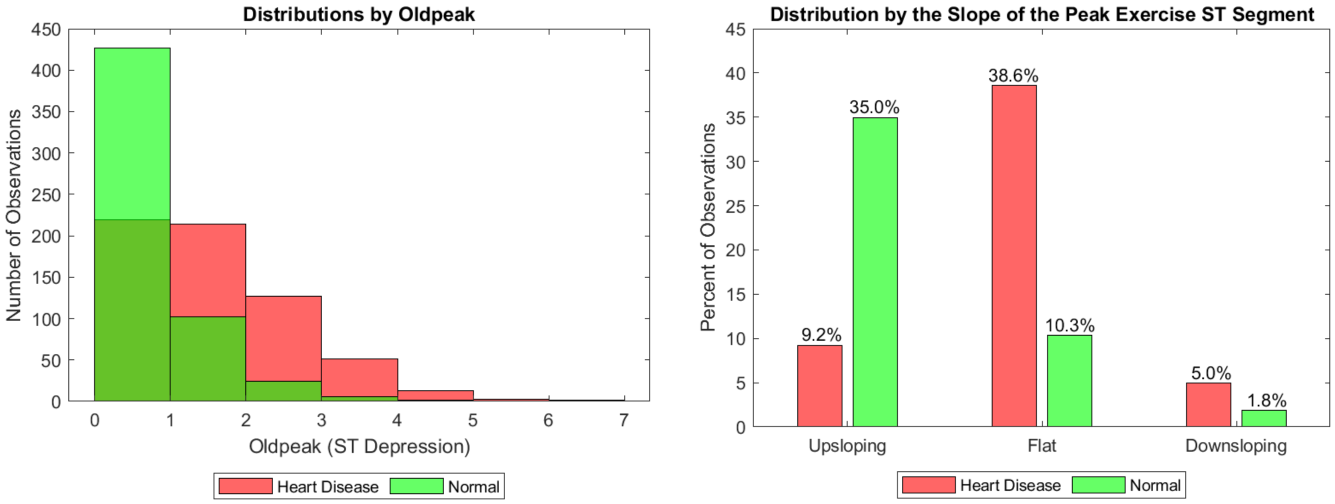

Fig. 6 Distribution by oldpeak (left) and the slope of the peak exercise ST segment (right)

D. *Classification*

After data preprocessing and examination, we divided the dataset into a training set (90%) and a test set (10%). Next, we trained and tested 43 machine learning classification models on the given dataset. The hyperparameters of 9 of the models were optimized using Bayesian optimization. We used the Classification Learner app [11] of the MATLAB R2024a [10] software for training, testing, and optimizing classification models. To prevent overfitting during the training, 10-fold cross-validation was used.

III. RESULTS

Table 1 shows the results of the training and testing. The best results were achieved using Ensemble and KNN machine learning classification models. Model #1 (Custom Ensemble) 's validation accuracy was 92.62%, and the same model's test accuracy was 95.80%. Model #2 (Custom KNN) reached 92.62% validation accuracy and 94.96% test accuracy.

Table 1. Accuracy, precision, recall, and f1 score of the compared classification models (sorted by validation accuracy)

#	Model Type	Preset	Validation (1071 observations)				Test (119 observations)			
			Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
1.	Ensemble	Custom Ensemble *	0.9262	0.9266	0.9347	0.9306	0.9580	0.9254	1.0000	0.9612
2.	KNN	Custom KNN *	0.9262	0.9178	0.9453	0.9314	0.9496	0.9118	1.0000	0.9538
3.	KNN	Weighted KNN	0.9150	0.9132	0.9277	0.9204	0.9580	0.9385	0.9839	0.9606
4.	Ensemble	Bagged Trees	0.9038	0.9143	0.9030	0.9086	0.9328	0.9091	0.9677	0.9375
5.	Neural Network	Wide Neural Network	0.8982	0.9119	0.8942	0.9029	0.9160	0.9194	0.9194	0.9194
6.	KNN	Fine KNN	0.8880	0.9042	0.8818	0.8929	0.9160	0.9194	0.9194	0.9194
7.	Ensemble	Boosted Trees	0.8833	0.8824	0.8995	0.8908	0.8067	0.7910	0.8548	0.8217
8.	SVM	Custom SVM *	0.8814	0.8873	0.8889	0.8881	0.8403	0.8209	0.8871	0.8527
9.	Neural Network	Medium Neural Network	0.8786	0.8827	0.8889	0.8858	0.8992	0.8788	0.9355	0.9063
10.	Neural Network	Trilayered Neural Network	0.8758	0.9033	0.8571	0.8796	0.9076	0.8696	0.9677	0.9160
11.	Tree	Fine Tree	0.8739	0.8776	0.8854	0.8815	0.8403	0.7945	0.9355	0.8593
12.	Tree	Custom Tree *	0.8739	0.8776	0.8854	0.8815	0.8403	0.7945	0.9355	0.8593
13.	SVM	Cubic SVM	0.8674	0.8801	0.8677	0.8739	0.8908	0.8769	0.9194	0.8976
14.	SVM	Medium Gaussian SVM	0.8655	0.8543	0.8995	0.8763	0.8319	0.8182	0.8710	0.8438
15.	Neural Network	Bilayered Neural Network	0.8637	0.8834	0.8554	0.8692	0.9328	0.9219	0.9516	0.9365
16.	Neural Network	Narrow Neural Network	0.8637	0.8636	0.8818	0.8726	0.8487	0.8143	0.9194	0.8636
17.	Ensemble	Subspace KNN	0.8590	0.8636	0.8713	0.8674	0.9160	0.8939	0.9516	0.9219
18.	Kernel	Logistic Regression Kernel	0.8562	0.8413	0.8977	0.8686	0.8151	0.7778	0.9032	0.8358
19.	SVM	Fine Gaussian SVM	0.8553	0.7934	0.9824	0.8779	0.8655	0.7949	1.0000	0.8857
20.	KNN	Medium KNN	0.8497	0.8651	0.8483	0.8566	0.7983	0.7879	0.8387	0.8125
21.	SVM	Quadratic SVM	0.8487	0.8497	0.8677	0.8586	0.8151	0.7941	0.8710	0.8308
22.	Kernel	SVM Kernel	0.8487	0.8358	0.8889	0.8615	0.8067	0.7910	0.8548	0.8217
23.	Kernel	Custom Kernel *	0.8478	0.8311	0.8942	0.8615	0.7983	0.7879	0.8387	0.8125
24.	Ensemble	RUSBoosted Trees	0.8459	0.8502	0.8607	0.8554	0.7731	0.7612	0.8226	0.7907
25.	KNN	Cubic KNN	0.8422	0.8566	0.8430	0.8498	0.8067	0.8197	0.8065	0.8130
26.	KNN	Cosine KNN	0.8413	0.8656	0.8289	0.8468	0.7983	0.8065	0.8065	0.8065
27.	Tree	Medium Tree	0.8394	0.8353	0.8677	0.8512	0.7647	0.7429	0.8387	0.7879
28.	SVM	Linear SVM	0.8357	0.8388	0.8536	0.8462	0.8235	0.8154	0.8548	0.8346
29.	SVM	Coarse Gaussian SVM	0.8338	0.8371	0.8519	0.8444	0.8235	0.8154	0.8548	0.8346
30.	Efficient Linear	Custom Efficient Linear *	0.8338	0.8418	0.8448	0.8433	0.8235	0.8154	0.8548	0.8346
31.	Discriminant	Linear Discriminant	0.8338	0.8394	0.8483	0.8439	0.8151	0.8125	0.8387	0.8254
32.	Discriminant	Quadratic Discriminant	0.8338	0.8467	0.8377	0.8422	0.8067	0.7910	0.8548	0.8217
33.	Discriminant	Custom Discriminant *	0.8338	0.8467	0.8377	0.8422	0.8067	0.7910	0.8548	0.8217
34.	Neural Network	Custom Neural Network *	0.8338	0.8455	0.8395	0.8425	0.7983	0.7879	0.8387	0.8125
35.	Binary GLM Logistic Regression	Binary GLM Logistic Regression	0.8319	0.8401	0.8430	0.8415	0.8235	0.8154	0.8548	0.8346
36.	Ensemble	Subspace Discriminant	0.8282	0.8377	0.8377	0.8377	0.8151	0.8030	0.8548	0.8281
37.	KNN	Coarse KNN	0.8282	0.8514	0.8183	0.8345	0.7983	0.7969	0.8226	0.8095
38.	Naive Bayes	Gaussian Naive Bayes	0.8245	0.8490	0.8131	0.8306	0.8319	0.8387	0.8387	0.8387

39.	Naive Bayes	Custom Naive Bayes *	0.8245	0.8490	0.8131	0.8306	0.8319	0.8387	0.8387	0.8387
40.	Efficient Linear SVM	Efficient Linear SVM	0.8226	0.8301	0.8360	0.8330	0.8403	0.8308	0.8710	0.8504
41.	Naive Bayes	Kernel Naive Bayes	0.8179	0.8240	0.8342	0.8291	0.7815	0.7571	0.8548	0.8030
42.	Efficient Logistic Regression	Efficient Logistic Regression	0.8049	0.8140	0.8183	0.8162	0.7731	0.7778	0.7903	0.7840
43.	Tree	Coarse Tree	0.7862	0.7752	0.8395	0.8061	0.8151	0.8125	0.8387	0.8254

* Bayesian optimization was used to optimize the hyperparameters of the model.

Both machine learning classification models with the highest validation accuracy are custom models, and Bayesian optimization was used to optimize their hyperparameters. In the next step, we show their minimum error hyperparameters.

Table 2 shows the bestpoint (=minimum error) hyperparameters of model #1 (Custom Ensemble). The Observed minimum classification error was 0.073733.

Table 2. Bestpoint (=minimum error) hyperparameters of model #1: Ensemble model

Hyperparameter	**Value**
Ensemble method:	GentleBoost
Number of learners:	220
Learning rate:	0.081521
Maximum number of splits:	245

Table 3 shows the bestpoint (=minimum error) hyperparameters of model #2 (Custom KNN). The observed minimum classification error was 0.073768.

Table 3. Bestpoint (=minimum error) hyperparameters of model #2: KNN model

Hyperparameter	**Value**
Number of neighbors:	43
Distance metric:	Jaccard
Distance weight:	Inverse
Standardize data:	true

Next, we were curious about the most important of the 11 predictors. For this reason, we inspected the two best models and calculated the permutation importance of the features. Fig. 7 shows the results for the Optimized Ensemble (model #1) and Optimized KNN (model #2). The results show that the most important feature was the slope of the peak exercise ST segment (STSlope) in both classification models.

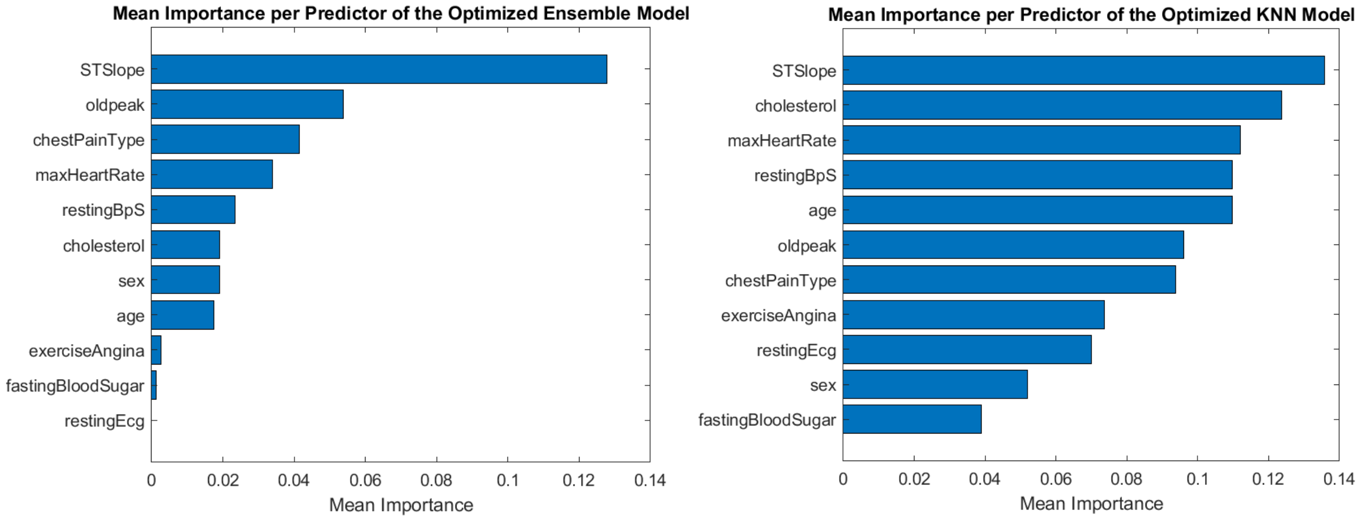

Fig. 7 Permutation importance of each feature in the classification models with the best validation accuracies

IV. DISCUSSION

As we can see in Table 1 of the results, some models reached high validation and test accuracy, but other models' accuracy was poor. It is essential to be aware that some models perform better on one dataset than others on another; the model's accuracy depends on the structure and data values of the given dataset. Therefore, several machine learning classification models on a given dataset should be tried to evaluate the various models' accuracy, precision, recall, and f1 score. For this reason, the Classification Learner App [11] of the MATLAB [10] software might be a helpful tool, as it relatively quickly compares the accuracy of the selected machine learning classification models on a preprocessed dataset. Afterward, it is possible to explore the chosen machine learning models more deeply, test them, and use them with new data.

On the heart disease dataset that we used [8][9], the best model has reached 92.62% validation accuracy and 95.80% test accuracy. We believe it is a good result; we can find similar accuracies for classification models on various cardiovascular datasets in the literature. Bhatt et al. [6] compared several classification models (such as decision tree, XGBoost, random forest, and multilayer perceptron) on a cardiovascular dataset of 70,000 instances and reached accuracies between 86–88%; the multilayer perceptron with cross-validation has outperformed all other algorithms. Subramani et al. [7] proposed a collection of machine learning models to predict cardiovascular disease. Their method provides nearly 96% accuracy results. The research of Ogunpola et al. [12] focused on detecting myocardial infarction using machine learning techniques, tackling the challenge of imbalanced datasets. They used seven machine learning and deep learning classifiers. The best accuracy, 98.50%, was reached using an optimized XGBoost model. Garg et al. [13] used KNN and random forest classification algorithms on a heart disease dataset; they obtained 86.885% accuracy for KNN and 81.967% for the random forest algorithm. Akkaya et al. [14] compared eight machine learning classification methods on a heart disease dataset. They treated the imbalance in the data, as well, with the production of synthetic data. The results showed that the XGBoost and KNN algorithms achieved the best accuracy; the XGBoost model reached 89% accuracy on the non-outlier data and 84.6% accuracy on the outlier data, and the KNN model reached 85.6% on the non-outlier data and 81% on the outlier data.

After calculating permutation feature importance on our first and second models, we have found that the most important predictor on the Optimized Ensemble (model #1) was the slope of the peak exercise ST segment, followed by oldpeak, chest pain type, maximum heart rate achieved, and resting blood pressure. In our second-best model (Optimized KNN), the most important predictor was the peak exercise ST segment, followed by serum cholesterol, maximum heart rate achieved, resting blood pressure, and age. The literature shows cardiovascular disease symptoms are usually chest pain, shortness of breath, poor blood supply to extremities, and fast or irregular heartbeat [15][16]. Risk factors for heart disease include age (growing older increases the risk), sex (men are at greater risk), family history, smoking, unhealthy diet (fat, salt, sugar), high blood pressure, high cholesterol, diabetes, obesity, lack of exercises, stress, or poor dental health [16]. As we can see, many of these symptoms and risks are related to the most important predictors of the Optimized Ensemble and Optimized KNN classification models.

V. CONCLUSION

In this research, we compared 34 classification models on a heart disease dataset [8][9] using MATLAB [10]. The dataset was divided into a training set (90%) and a test set (10%). During training, 10-fold cross-validation was used. The best results were reached with the Custom Ensemble classification model (92.62% validation accuracy and 95.80% test accuracy) and the Custom KNN classification model (92.62% validation accuracy and 94.96% test accuracy). During the training of these models, Bayesian optimization was used to optimize the model’s hyperparameters. Finally, we used permutation importance to detect which features were most important. The results showed that the slope of the peak exercise ST segment was the most important predictor in both classification models.

The results and methodology of this research can be used to utilize machine learning classification models for diagnosing cardiovascular disease and, in future research, to analyze, compare, and explore several classification models on various datasets.

REFERENCES

[1] World Heart Federation (2023) World Heart Report 2023. Confronting The World’s Number One Killer. [Online]. Available: https://world-heart-federation.org/wp-content/uploads/World-Heart-Report-2023.pdf

[2] (2022) Cardiovascular Disease. [Online]. Available: https://my.clevelandclinic.org/health/diseases/21493-cardiovascular-disease

[3] L. Végh, K. Czakóová, and O. Takáč, “Comparing Machine Learning Classification Models on a Loan Approval Prediction Dataset,” *International Journal of Advanced Natural Sciences and Engineering Researches*, vol. 7, no. 9, 2023, pp. 98–103. https://doi.org/10.59287/ijanser.1516

[4] M. T. Fülöp, M. Gubán, Á. Gubán, and M. Avornicului, “Application Research of Soft Computing Based on Machine Learning Production Scheduling,” Processes, vol. 10. no. 3, 2022, paper 520. https://doi.org/10.3390/pr10030520

[5] J. Udvaros and N. Forman, “The Merger of Machine Learning and Artificial Intelligence: New Horizons in Education 4.0,” in *ICEBM 2023 6th International Conference on Economics and Business Management*, Cluj-Napoca, Romania, 2023, p. 48.

[6] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, “Effective Heart Disease Prediction Using Machine Learning Techniques,” *Algorithms*, vol. 16, issue 2, 2023, paper 88. https://doi.org/10.3390/a16020088

[7] S. Subramani, N. Varshney, M. V. Anand, M. E. M. Soudagar, L. A. Al-keridis, T. K. Upadhyay, N. Alshammari, M. Saeed, K. Subramanian, K. Anbarasu, and K. Rohini, “Cardiovascular diseases prediction by machine learning incorporation with deep learning,” *Frontiers in Medicine*, vol. 10, 2023, paper 1150933. https://doi.org/10.3389/fmed.2023.1150933

[8] M. Siddhartha. (2024) Heart Disease Dataset. [Online]. Available: https://www.kaggle.com/datasets/mexwell/heart-disease-dataset/data

[9] M. Siddhartha. (2020) Heart Disease Dataset (Comprehensive). [Online] IEEE Dataport. Available: https://dx.doi.org/10.21227/dz4t-cm36

[10] (2024). MATLAB. [Online]. Available: https://www.mathworks.com/products/matlab.html

[11] (2024). Classification Learner. [Online] Available: https://www.mathworks.com/help/stats/classificationlearner-app.html

[12] A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarrak, and S. N. Qasem, “Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases,” *Diagnostics*, vol. 14, no. 2, 2024, paper 144. https://doi.org/10.3390/diagnostics14020144

[13] A. Garg, B. Sharma, and R. Khan, "Heart disease prediction using machine learning techniques," in *IOP Conference Series: Materials Science and Engineering*, vol. 1022, 2021, paper 012046. https://doi.org/10.1088/1757-899X/1022/1/012046

[14] B. Akkaya, E. Sener, and C. Gursu, "A Comparative Study of Heart Disease Prediction Using Machine Learning Techniques," in *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, Ankara, Turkey, 2022, pp. 1–8. https://doi.org/10.1109/HORA55278.2022.9799978

[15] (2024) Warning signs and symptoms of heart disease. [Online] Available: https://www.mountsinai.org/health-library/selfcare-instructions/warning-signs-and-symptoms-of-heart-disease

[16] (2024) Heart disease. [Online] Available: https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118