The Classification of Coronary Artery Disease Using A Machine Learning Approach: A Preliminary Study

Ahyar Supani[1,2]*, Siti Nurmaini[3], Radiyati Umi Partan[4] and Bhakti Yudo Suprapto[5]

[1] Doctoral Program of Engineering Science, Universitas Sriwijaya, Palembang, Indonesia
[2] Polytechnic State of Sriwijaya, Palembang, Indonesia
[3]Intelligent System Research Group, Universitas Sriwijaya, Indonesia
[4] Interrnal Medicine Department, Medicine Faculty, Universitas Sriwijaya, Indonesia
[5] Electrical Engineering Department, Universitas Srwijaya, Indonesia

ahyarsupani@polsri.ac.id

Abstract. Coronary heart disease (CAD) is the world's leading cause of death. Early detection of this condition is critical. Diagnosis with visual images through examination with angiography techniques is the current gold standard. However, this technique causes side effects, so it is necessary to carry out alternative examinations based on symptoms that do not pose a risk. To increase the accuracy of CAD diagnosis based on symptoms, it can use computer assistance via machine learning. The aim of this research is to classify CAD or normal patients using a machine learning model approach using the Naïve Bayes, XGboost, and K-Nearest Neighbor (KNN) algorithms. The dataset used in the experiment is Z-Alizadeh Sani, which contains 55 features. The experimental results were evaluated using the metrics accuracy, IoU, precision, recall, and F1_score. Of the three algorithms tested, XGBoost produced the four highest scores regarding the metrics accuracy (0.852), IoU (0.824), recall (0.977), and F1_score (0.903), which outperformed the other two algorithms. KNN had the highest metric precision, with a value of 0.870. Overall, XGBoost remains superior in performance.

Keywords: Xgboost, KNN, Machine Learning, Coronary Heart Disease.

1 Introduction

Coronary disease is one of the heart diseases that is the main cause of death in the world, with an estimated 7.3 million people dying out of the 17.3 million who die due to CVD [1]. CAD occurs due to narrowing of the three main coronary blood vessels, which include (i) the left circumflex artery (LCX), (ii) the right coronary artery (RCA), and (iii) the left anterior descending [LAD] [2, 3, 4, 5]. Narrowing is caused by atherosclerotic plaques, for example, mixed atherosclerotic plaques [6, 7], so detection and analysis of multi-class atherosclerotic plaques are very important for prevention and early examination [8].

CAD examination can be carried out by diagnosis through several tests, namely (i) angiography, (ii) echocardiogram, (iii) electrocardiogram (ECG), (iv) exercise stress test, and (v) electron beam (ultrafast) CT scans [9]. Examination using the

N. L. Husni et al. (eds.), *Proceedings of the 7th FIRST 2023 International Conference on Global Innovations (FIRST-ESCSI 2023)*, Advances in Engineering Research 232,
https://doi.org/10.2991/978-94-6463-386-3_2

angiography method is the gold standard and is trusted by doctors, but this method is invasive, expensive, and causes side effects such as arrhythmia, arterial dissection, and even death [1]. So alternative examinations are needed based on symptoms that do not pose a risk. An examination based on symptoms by a doctor requires an accurate diagnosis. This can be done with the help of a computer through a machine learning model approach.

The application of the machine learning model includes several algorithms, namely SVM, Random Forets, Decision Tree, Naïve Bayes, K-NN, and xGBoost. Among these models, xGBoost has the highest accuracy, according to several studies. In research [10], XGBoost is applied in the interpretation and application of post-fault transient stability status prediction for power systems. Li et al. [11] predicted very high gene expression values in humans using XGBoost. Hendrawan et al. [12] compared the XGBoost algorithm with Naïve Bayes to classify local product review text. Ramraj et al. [13] classify and predict different datasets with XGBoost. Zhang et al. [14] improved the performance of the xGBoost model for CAD prediction and feature processing. Budholiya et al. [15] predicted heart disease with an optimized XGBoost. Pan et al. [16] applied an optimized XGBoost to predict reservoir porosity using petrophysical logs. Pathan et al. [17] predicted the accuracy of heart disease using feature selection.

A decision tree ensemble based on gradient boosting with a high degree of scalability is called XGBoost. XGBoost constructs an additive expansion of the objective function by minimizing the loss function, much to gradient boosting. Given that XGBoost only uses decision trees as a fundamental classifier, the complexity of the tree is managed using a variety of loss functions [18, 19, 20].

Apart from that, according to [1], the Z-Alizadeh Sani dataset is an up-to-date CAD dataset and is widely used by researchers to create models in terms of CAD classification. This dataset has 54 features as input, 1 feature as output, and 303 patients. In the initial study, this paper proposed three algorithms, namely XGBoost, k-Nearest Neighbor (K-NN), and Naïve-Bayes for machine learning models that are useful for classifying CAD patients. The dataset used in this paper is Z-Alizadeh Sani. Two algorithms, KNN and Naive-Bayes, are used to compare the XGBoost algorithm with the aim of analyzing its performance against XgBoost. The research objectives are (i) to classify CAD patients using three machine learning models as a preliminary study, and (ii) to compare the three models to obtain high performance as the best model. The contributions of this paper are (i) obtaining three machine learning models for CAD classification; (ii) comparing the three models to get the highest and best performance; (iii) applying the three machine learning models to a CAD dataset consisting of fifty-five features; (iv) evaluating algorithms Naïve Bayes, XGBoost, and KNN to diagnose CAD.

This paper is structured as follows: section 1 explains the introduction, section 2 explains the methodology, section 3 explains the results and discussion, and section 4 concludes.

2 Methods

2.1 Data Acquisition

The data acquisition process was carried out by downloading the coronary heart disease (CAD) dataset originating from the UCI and Kaggle dataset locations with addresses: (https://www.kaggle.com/datasets/tanyachi99/zalizadeh-sani-dataset-2csv). The data taken consisted of 303 data points on coronary heart disease (CAD) and healthy patients, where each data point had 55 variables, consisting of 54 features used in the test dataset as independent variables, namely **Age** states the patient's age, **Weight** states the patient's weight, **Length** states the patient's height, **Sex** states the patient's gender, and **body mass index** (BMI) states the patient's body mass based on weight and height. **Diabetes mellitus** (DM) states whether the patient has diabetes or not. **Hypertension** (HTN) states the patient's blood pressure history, **Current Smoker** states whether the patient is currently smoking or not; **Ex-Smoker** states whether the patient has previously smoked or not; **Family history** (FH) describes a history of heart disease; **obesity** states that the patient is overweight, **Chronic Renal Failure** (CRF), **Cerebrovascular Accident** (CVA), **Airway disease**, **Thyroid Disease**, **Congestive Heart Failure** (CHF), **Dyslipidemia** (DLP), **BP**, **PR, Edema, Weak Peripheral Pulse, Lung rales, Systolic Murmur, Diastolic Murmur, Typical Chest Pain** states whether the patient feels chest pain or not; **Dyspnea, Function Class, Atypical, Nonanginal, Exertional CP, LowTHAng, Q Wave, St Elevation, St Depression, Tinversion, LVH, Poor R Progression, FBS, CR, TG, Low density lipoprotein (**LDL**), High density lipoprotein (**HDL**), Blood Urea Nitrogen (**BUN**), Erythrocyte Sedimentation rate (**ESR**), Hemoglobin (**Hb**), Potassium (**K**), Sodium (**Na**), White Blood Cell (**WBC**), Lymphocyte (**Lymph**), Neutrophil (**Neut**), Platelet (**PLT**), Ejection Fraction (**EF-TTE**), Regional Wall Motion Abnormality (**RWMA**),** and **Valvular Heart Disease (**VHD**).** A feature is a target, namely **The class attribute** (Cath) states the patient's condition is CAD or normal, where CAD patients have a value of '0', and normal has a value of '1'

2.2 Algorithms

1. Naïve Bayes

One of the most successful and efficient inductive learning algorithms for machine learning and data mining is Nave Bayes. Despite using the concept of attribute independence (no association between attributes), Nave Bayes performs competitively in the classification process. In actual data, the assumption of independence of these qualities is rarely violated, but even if the assumption of independence of these attributes is breached, the performance of the naive Bayes classifier is relatively high, as demonstrated by many empirical research. The Bayes' theorem is a categorization that uses probability and statistical techniques to forecast future opportunities based on historical data. This theorem is paired with "naive," which makes the erroneous assumption that the relationships between qualities are independent of one another. Each row or document I in a dataset is taken to be a vector of attribute values, where each value represents an attribute review, Xi (i[1,n]). Each row has a class label ci Є {c1,c2,…,ck}as the value of the class C variable, so that to carry out classification, the probability value can be calculated p(C=ci|X=xj). Because in Naïve Bayes it is assumed that each attribute is independent, the equation obtained is as follows [21, 22]:

$$P\left(label = c_j IY\right) = \frac{P(YIlabel=c_j)*P(c_j)}{P(Y)} \quad (1)$$

The dominator P (Y) can be eliminated safely because it is independent of labels. The class label cj has the largest conditional probability value, determining the category of the data record.

2. xGBoost

XGBoost is an open-source software package that provides a regularizing gradient boosting framework for C++, Java, Python, R, Julia, Perl, and Scala. It is compatible with Linux, Windows, and macOS. While, A machine learning method called gradient boosting is used, among other things, for classification and regression tasks. It provides a prediction model in the form of an ensemble of weak prediction models, i.e., models that make only a few data-related assumptions and are frequently straightforward decision trees. The resulting technique, known as gradient-boosted trees, typically beats random forest when a decision tree is the weak learner. The construction of a gradient-boosted trees model follows the same stage-wise process as previous boosting techniques, but it generalizes other techniques by enabling optimization of any differentiable loss function [10, 18, 20].

3. K-Nearest Neighbor (KNN)

The "Supervised Learning Algorithm" includes the complex classification technique known as the "K-Nearest Neighbour Algorithm," which is utilized for regression and categorization. It is a flexible and adaptable method that can also be applied to assessments with missing data and datasets with examples. It uses K-Nearest Neighbors Data objects, as the name suggests, to predict and identify the category or continuous assessment for the creative original data item. It is categorized as a lax prediction technique. This algorithm sorts the training data, and the separation between training samples and instances is computed. All categories can have defined training data. Most of the nearest neighbors of the query record have a predicted value. [23].

$$d(a,b) = \sqrt{\sum_{i=1}^{n} (a_i - b_j)^2} \tag{2}$$

The two different data objects are represented by parameters a and b in Equation (2).

2.3 Metrics evaluation

In this paper, we propose several evaluation metrics, namely accuracy, IoU, F1-score, precision, and recall [24, 25], which are described in the confusion report. Equation (3-7) for this metric is

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \tag{3}$$

$$IoU = \frac{Intersection}{Union} = \frac{TP}{TP+FN+FP} \tag{4}$$

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

$$F_1 score = \frac{2\ TP}{2\ TP+FP+FN} \tag{7}$$

Where,

TP= True Positive; TN = True Negative; FP = False Positive; FN = False Negative.

3 Results

In this paper, we have tried to train features for the causes of coronary heart disease with three algorithms. In this section, we explain the test results for each algorithm trained. Each trial of the Naïve Bayes, xGBoost, and K-NN algorithms has been preprocessed with standard type scaling before being processed by the model, specifically for xGBoost and K-NN, except Naïve Bayes. Model experiments have implemented model optimization with grid search as model tuning and cross-validation as a division of training and validation datasets. Tuning the model and preprocessing this dataset aim to improve model performance. Algorithm performance evaluation is carried out using accuracy, IoU, precision, recall, and F1_score metrics. Table 1 shows the accuracy performance results for the three algorithms tested, which include accuracy score model training, the best score model, and the score testing model. The xGBoost algorithm outperforms the other two algorithms, which have an accuracy model score of 0.852.

Table 1. Accuracy score (model training, model best, and model testing) for three models

Type of algorithms	Accuracy			remarks
	Model score training	Model best score	Model score testing	
Naïve Bayes	0.752	0.678	0.689	without scaling
xGBoost	**1**	**0.888**	**0.852**	scaling
KNN	0.873	0.873	0.846	scaling

Then, Fig. 1 is a multinomial Naïve Bayes confusion matrix that has parameter values in the test score: TP = 31, TN = 11, FP = 7, and FN = 12.

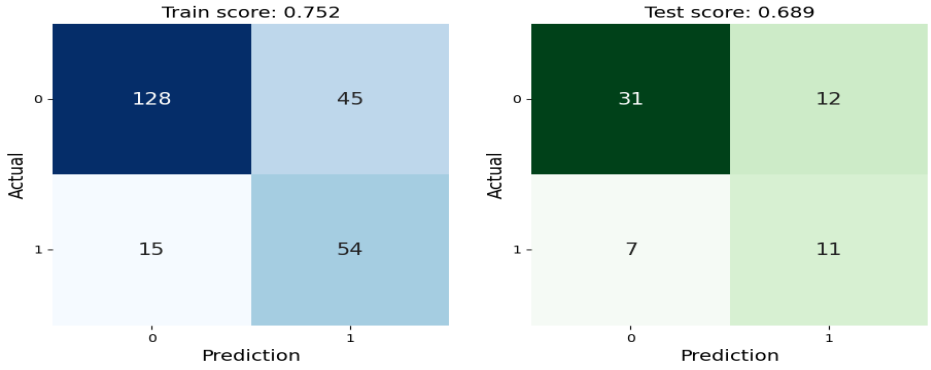

Fig 1. Confusion matrix for multinomial NaiveBayes.

Fig. 2 shows the confusion matrix for xGBoost with test score parameter values TP = 42, TN = 10, FP = 8, and FN = 1. This algorithm produces an accuracy of 1 during training and an accuracy value of 0.852 during testing.

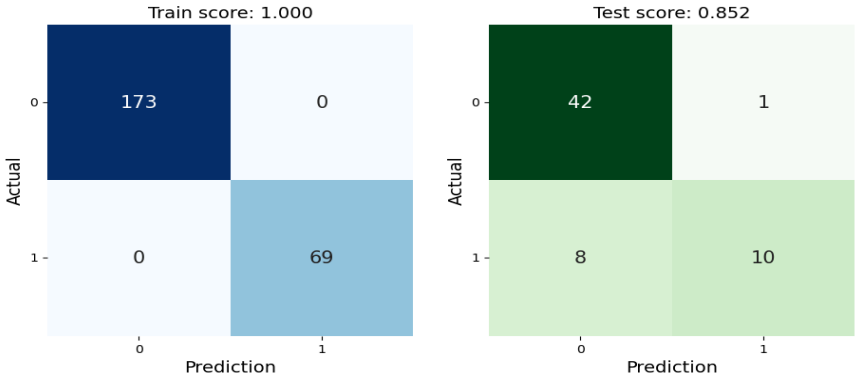

Fig.2. Confusion matrix for XGBoost

Fig. 3 shows the confusion matrix for the KNN algorithm with test score parameter values TP = 60, TN = 17, FP = 9, and FN = 5. This model produces an accuracy score of 0.873 during training and an accuracy score of 0.846 during testing.

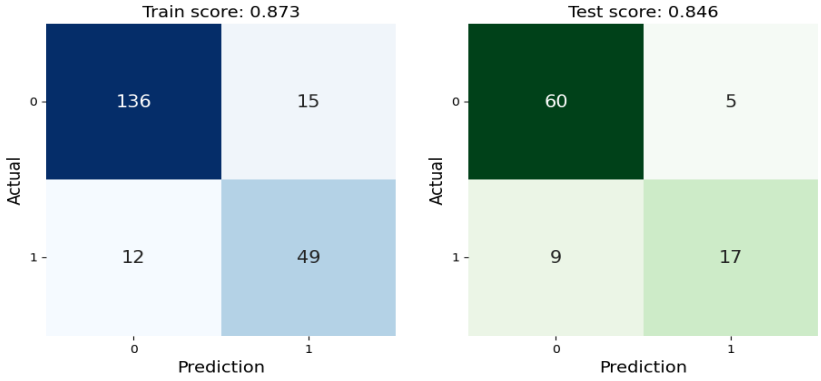

Fig. 3. Confusion matrix for KNN

Next, calculate the other metric values: IoU precision, recall, and F1_score manually using equations 4–7. The metric values are obtained according to Table 4 below.

Table 2. Evaluation of five metrics for three algorithms

	Accuracy	IoU	Precision	Recall	F1-score

Naïve Bayes	0.689	0.620	0.816	0.721	0.765
xGBoost	**0.852**	**0.824**	0.840	**0.977**	**0.903**
KNN	0.846	0.811	**0.870**	0.923	0.896

4 Discussions

The experimental results of three algorithms in Table 2 show that the best model is xGBoost for four metrics except precision. This model approaches almost perfect prediction results with test data. Testing method with 3-fold cross-validation, where the test data is never randomized or combined with the training data. Then the training data is carried out in a stratified manner each time the machine is trained, so that normal and sick data are always representative during training.

Table 5 shows a comparison with other methods in terms of classification. Zhang et al. [14] classified CAD patients with the Z-Alizadeh Sani dataset with the XGBoost algorithm. They applied feature selection to the dataset and cross-validation 10, and then performed CAD classification, which resulted in a score of 0.894. Their model outperforms our model, where our proposed model has not implemented feature selection but has implemented cross-validation 3. Ramraj [13] applied the XGBoost method for the classification of diabetes patients. The results obtained were 0.767 for accuracy. Our proposed model outperforms their model. Then Gupta et al. [26] classified cardiac disease patients with a dataset that had 14 features. Their model is slightly superior to our model, but their model has a dataset of 14 features, which is different from the dataset we used with 55 features. Although our proposed results are close to superior to previous research, our method does not yet perform feature selection.

Table 3. Comparison of other methods for classification

Authors	Method	Accuracy
Zhang et al. [14]	XGBoost, feature selection	0.894
Ramraj et al. [13]	XGBoost	0.767
Gupta et al. [26]	Random Forest	0.857
Proposed method	XGBoost	**0.852**

For further research, we will carry out feature selection or feature importance to improve the performance of our model.

5 Conclusions

In the preliminary study, of the three models that we applied for CAD classification with a dataset of 55 features and not yet performing feature selection. XGBoost outperforms Naïve Bayes and K-Nearest Neighbor in terms of accuracy, IoU, recall, and F1_score, which are 0.852, 0.824, 0.977, and 0.903, respectively. These results are good for using CAD diagnosis with the help of machine learning in clinical practice.

References

1. Shahid, Afzal & Singh, Maheshwari & Roy, Bishwajit & Aadarsh, Aashish : Coronary Artery Disease Diagnosis Using Feature Selection Based Hybrid Extreme Learning Machine. In: 3rd Internatioal Conference on Information and Computer Technologies (ICICT), pp. 341-346 (2020).
1. M. Abdar, W. Książek, U. R. Acharya, R. S. Tan, V. Makarenkov, and P. Pławiak, :A new machine learning technique for an accurate diagnosis of coronary artery disease. Comput. Methods Programs Biomed (179), (2019).
2. M. A. Gatzoulis, S. Y. Ho, and E. D. Nicol : Great Vessel and Coronary Artery Anatomy in Transposition and Other Coronary Anomalies: A Universal Descriptive and Alphanumerical Sequential Classification. JACC: Cardiovascular Imaging 6(5), (2013).
3. M. Darvishi and A. Moayeri : Anatomical indicators of the heart and coronary arteries: An anthropometric study. Biomed. Res. Ther. 7(9), 3977–3984 (2020).
4. G. Thiene, C. Frescura, M. Padalino, C. Basso, and S. Rizzo : Coronary arteries: Normal anatomy with historical notes and embryology of main stems. Front. Cardiovasc. Med. 8, 1–12 (2021).
5. A. I. Guaricci et al. : The presence of remodeled and mixed atherosclerotic plaques at coronary ct angiography predicts major cardiac adverse events—the CAFE-PIE Study. Int. J. Cardiol. 215, 325-331 (2016).
6. U. R. Acharya et al.: Atherosclerotic plaque tissue characterization in 2D ultrasound longitudinal carotid scans for automated classification: A paradigm for stroke risk assessment. Med. Biol. Eng. Comput. 51(5), 513–523 (2013).
7. F. Zhao et al. : An automatic multi-class coronary atherosclerosis plaque detection and classification framework. Med. Biol. Eng. Comput. 57(1), 245–257 (2018).
8. Tan, J.H., et al. : Application of stacked convolutional and long shortterm memory network for accurate identification of CAD ECG signals. Computers in biology and medicine 94, 19-26 (2018).
9. Minghua Chen, Qunying Liu, Yicen Liu, Chang-Hua Zhang, Ruihua Liu : XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System. IEEE Acces 7, 2169-3536 (2019).
10. Wei Li, Yanbin Yin, Xiongwen Quan and Han Zhang. Gene expression value prediction based on XGboost algorithm. Frontiers in Genetics 10, (2019).
11. Ivan Rifky Hendrawan, Ema Utami, Anggit Dwi Hartanto : Comparison of Naïve Bayes Algorithm and XGBoost on Local Product Review Text Classification. Edumatic: Jurnal Pendidikan Informatika 6(1), 143-149 (2022).
12. Ramraj S, Nishant Uzir, Sunil R and Shatadeep Banerje : Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets. International Journal of Control Theory and Applications 9(40), 651-662 (2016).
13. Shasha Zhang, Yuyu Yuan, Zhonghua Yao , Xinyan Wang and Zhen Lei : Improvement of the Performance of Models for Predicting Coronary Artery Disease Based on XGBoost Algorithm and Feature Processing Technology. Electronics 11(315) , (2022).
14. Kartik Budholiya, Shailendra Kumar Shrivastava, Vivek Sharma. An optimized XGBoost based diagnostic system for effective prediction of heart disease. Journal of King Saud University–Computer and Information Sciences 34, 4514–4523 (2022).
15. Shaowei Pan, Zechen Zheng, Zhi Guo, Haining Luo : An optimized XGBoost method for predicting reservoir porosity using petrophysical logs. Journal of Petroleum Science and Engineering 208, 109520 (2022).
16. Muhammad Salman Pathan, Avishek Nag, Muhammad Mohisn Pathan, Soumyabrata Dev : Analyzing the impact of feature selection on the accuracy of heart disease prediction. Healthcare Analytics 2, 100060 (2022).

17. Tianqi Chen, Carlos Guestrin : XGBoost: A Scalable Tree Boosting System. arXiv:1603.02754v3 (2016).
18. Baris Balaban, Caglar Yilgor, Altug Yucekul, Tais Zulemyan, Ibrahim Obeid, Javier Pizones, Frank Kleinstueck, Francisco Javier Sanchez Perez-Grueso, Ferran Pellise, Ahmet Alanay, Osman Ugur Sezerman : Corrigendum to "Building clinically actionable models for predicting mechanical complications in postoperatively well-aligned adult spinal deformity patients using XGBoost algorithm. Inform Med Unlocked 37, 101191 (2023).
19. Candice Bent ejac, Anna Csorgo, Gonzalo Mart ınez-Mu noz. A Comparative Analysis of XGBoost. arXiv:1911.01914 (2019).
20. Syarli, Asrul Ashari Muin : Metode Naive Bayes Untuk Prediksi Kelulusan. Jurnal Ilmiah Ilmu Komputer 2(1), (2016).
21. R. Alizadehsani, Jafar Habibi, Mohammad Javad Hosseini, Reihane Boghrati, Asma Ghandeharioun, Behdad Bahadorian, Zahra Alizadeh Sani : Diagnosis of Coronary Artery Disease Using Data Mining Techniques Based on Symptoms and ECG Features. European Journal of Scientific Research 82(4), 542-553 (2012).
22. C. A. Bhardwaj, M. Mishra, and K. Desikan : Dynamic feature scaling for K-nearest neighbor algorithm. https://arxiv.org/abs/1811.05062v1. (2018).
23. Abdel Aziz Taha and Allan Hanbury : Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Medical Imaging 15(29), (2015).
24. Bilel Benjdira, Adel Ammar, Anis Koubaa, and Kais Ouni : Data-Efficient Domain Adaptation for Semantic Segmentation of Aerial Imagery Using Generative Adversarial Networks. Appl. Sci. 10(1092), (2020).
25. Chiradeep Gupta, Athina Saha, N V Subba Reddy, U Dinesh Acharya : Cardiac Disease Prediction using Supervised Machine Learning Techniques. Journal of Physics: Conference Series 2161 (012013), (2022).