Scientific Journal of Informatics
Vol. 11, No. 2, May 2024

p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Comparison Model Optimal Machine Learning Model With Feature Extraction for Heart Attack Disease Classification

Salsa Desmalia[1*], Amril Mutoi Siregar[2], Kiki Ahmad Baihaqi[3], Tatang Rohana[4]
[1,2,3,4] Department of Informatics, Faculty of Computer Science, Buana Perjuangan University, Karawang, Indonesia

Abstract.
Purpose: The purpose of this study is to classify the number of people affected by heart disease and those not affected by heart disease based on various categories of heart attack causes. This study aims to urge people to take better care of their health and to serve as a reference for doctors to educate patients about the dangers of heart attacks.
Methods: The model will be constructed via a machine learning methodology. The algorithms utilized in its development encompass the Support Vector Machine (SVM) algorithm, the K-Nearest Neighbor (k-NN) algorithm, and the Random Forest (RF) algorithm. This study utilizes principal component analysis (PCA) as a means of extracting optimized features from the dataset, employing techniques for dimension reduction prior to modeling the data.
Result: Cumulative explication of the concept of variance constitutes a foundational aspect of PCA (principal component analysis) within the scope of the current research, namely a dimensionality reduction technique employed in multivariate data analysis to facilitate model development, thereby enabling the creation of more optimal and comprehensive models. In this research, the dimensions of training data are incorporated during the process of model creation. The results show KNN model exhibits the highest performance, with an accuracy of 86%, precision of 86%, recall of 91%, and F1-score of 88%. Furthermore, evaluation using the ROC metric also provides a relatively favorable value, 0.85.
Novelty: Researchers used 1190 patient data sourced from Kaggle. Before modeling the algorithm, researchers conducted EDA & Preprocessing which includes missing values to find data that does not have information, then duplicate data to find duplicated data, there are 270 duplicated data, then the duplicated data is deleted so that the data becomes 737, then PCA implementation is carried out. PCA is reducing features automatically without changing the data.

Keywords: Heart attack, Classification, Machine learning, PCA, ROC curve
Received May 2024 / **Revised** May 2024 / **Accepted** May 2024

INTRODUCTION

Heart disease is currently one of the major health problems that many people in the world have to deal with [1]. According to WHO, CVD/HD accounts for 17.9 million deaths worldwide each year, making it the leading cause of death. It covers a wide range of heart and vascular diseases, such as rheumatic HD, coronary HD, and cerebrovascular disease. [2]. Therefore, there are several body factors that must be considered such as blood pressure, cholesterol, diabetes, smoking, and gender, all of which are associated with heart disease patients [3].

Two types of causes of heart failure have been identified. The first is related to cardiac anomalies, which include conditions such as heart attacks. The second is associated with excessive blood pressure. Treatment of heart failure can be achieved through preventive measures, including lifestyle modification, medication, and surgery. Additionally, the diagnosis of heart failure can be facilitated through the use of artificial intelligence, particularly machine learning [4]. Machine learning is a subfield of artificial intelligence that deals with statistics, modifying algorithms to enhance the accuracy and efficiency of data analysis. It is particularly adept at developing acceptance policies and self-learning algorithms. Another area where machine learning excels is big data management. It is well-suited to the analysis of a wide variety of information, both labeled and unlabeled [5]. In order to accurately diagnose heart failure disease using various machine learning models on various data sets, substantial amount of research and development has

[*] Corresponding author.
Email addresses: if20.salsadesmalia@mhs.ubpkarawang.ac.id (Desmalia)
DOI: 10.15294/sji.v11i2.4561

been conducted. The use of a variety of data mining techniques and neural networks has permitted the determination of the degree of severity of human heart disease. This classification has been achieved through a combination of methods, including the Decision Tree (DT), K-Nearest Neighbor Algorithm (KNN), Naive Bayes (NB), and Genetic Algorithm (GA).

Some previous studies have successfully used these techniques to classify data usage [6], [7]. Then the research conducted by [8]. Classifying heart disease with employing machine learning approach can be interpreted in an ECG (Electrocardiogram), which is the least expensive or non-invasive diagnostic technique. In this research, several machine learning algorithms were employed, including AdaBoost, Random Forest, CNN, XGBoost, and Decision Tree [9] .This research uses a dataset whose variables consist of cholesterol, height, gender, age, and others. Using Support vector classifier algorithm, Random Forest, Decision Tree. This research classifies K-neighbor with 87% accuracy, while the accuracy of Support Vector, Decision Tree, and Random Forest are 83%, 79%, and 84%.

Based on the problems that have been described, this research aims to classify those affected by heart disease and those not affected by heart disease with various machine learning methods. What distinguishes this research from the previous one is in terms of data and different algorithms, before modeling the algorithm, researchers use Principal Component Analysis (PCA) to select the columns that have the most influence on each other. Researchers use the KNN, Random Forest and SVM algorithms using Confusion Matrix and ROC evaluation comparisons to improve and determine the accuracy results of the algorithms used in previous studies.

METHODS

Dataset

The present study employs a dataset sourced from the Kaggle (https://www.kaggle.com/datasets/sid321axn/heartstatlog-cleveland-hungary-final). It encompasses 1,190 rows with 12 attributes and two distinct classes. The 12 attributes consist of Age, Sex, Chest Pain Type, Resting bp s, Cholesterol, Blood Sugar, Resting ECG, Max Heart Rate, Exercise Angina, Oldpeak, ST Slope. Next, categorize with 2 classes, namely affected by heart disease and not affected by heart disease. Figure 1 is a snapshot of the dataset used.

Table 1**.** Kaggle dataset of Heart Disease Description

No	Feature of dataset	Description
1	Age	The age of the patient in years (Numeric)
2	Sex	The gender of the patient was recorded as male (1) or female (0) (Nominal)
3	Chest pain type	In terms of classification, the type of chest pain divided into four categories: typical, typical angina, non-anginal pain, and asymptomatic.
4	Resting bp s	The resting blood pressure level in millimeters of mercury (mm Hg) (Numerical)
5	Cholesterol	The serum cholesterol concentration is expressed in milligrams per deciliter (mg/dl) (Numeric)
6	Fasting blood sugar	A fasting blood sugar level of greater than 120 mg/dl is indicative of a true positive, while a fasting blood sugar level of less than 120 mg/dl is indicative of a false negative.
7	Resting ecg	The results of an electrocardiogram while at rest are represented by three distinct values: 0, indicating normal results; 1, indicating an abnormality.
8	Max heart rate	The maximum heart rate achieved during the exercise session was recorded (Numeric)
9	Exercise angina	Angina induced by exercise 0 depicting NO 1 depicting Yes (Nominal
10	Oldpeak	Exercise induced ST-depression in comparison with the state of rest (Numeric)
11	ST slope	ST segment measured in terms of slope during peak exercise 0: Normal 1: Upsloping 2: Flat 3: Downsloping
12	Target	A value of 1 on the Heart Risk scale indicates the presence of heart disease, while a value of 0 indicates the absence of heart disease

The research procedure starts with researchers looking for datasets and then analyzing the data set, the next step is to do EDA & preprocessing, the next step is to implement PCA to select the columns that have the most influence on each other without changing the data, then modeling algorithms with SVM, Random Forest, KNN, then evaluate using confusion matrix and ROC.

Research procedure

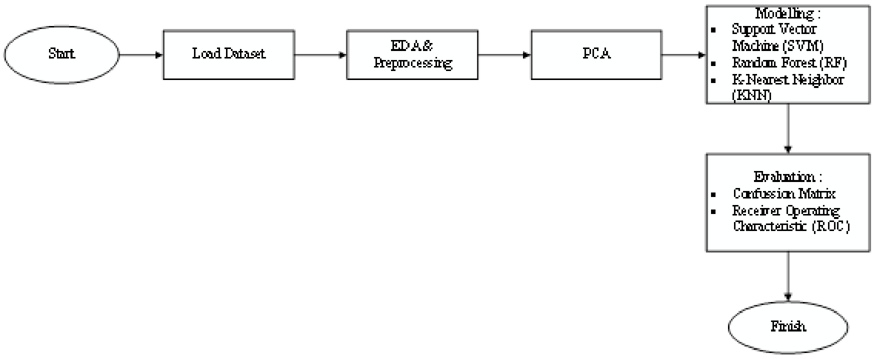


Figure 1. Flowchart of research procedure

The research procedure starts with researchers looking for datasets and then analyzing the data set, the next step is to do EDA & preprocessing, the next step is to implement PCA to select the columns that have the most influence on each other without changing the data, then modeling algorithms with SVM, Random Forest, KNN, then evaluate using confusion matrix and ROC.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a methodology employed to enhance the comprehension of a data set under investigation. This is achieved by employing statistical graphs and other data visualizations. EDA is comprised of four primary components. This encompasses measures of central tendency (such as the mean, mode, and median), measures of dispersion (such as standard deviation and variance), and measures of the distribution shape. It also involves the identification of potential outliers. Datasets usually include a variety of important information that is difficult to find. By using statistical analysis and EDA, the processed data can be used. EDA is used to analyze data before modeling [10].

Support Vector Machine (SVM)

Support Vector Machine which is a machine learning method used to solve various problems such as classification, regression, and other detection problems. Used for classifiers, the binary SVM decision boundary uses a hyperplane with maximum margin. By finding the point with the greatest distance between two information points from two categories which is the goal. [11]. In SVM, vectors are important in the classification process, a vector is an object that has a value of Vector Magnitude and Vector Direction. Mathematical vector calculation of classification in the SVM algorithm [12].

Random Forest

Random Forest is an extension of Decision Tree where training uses individual samples and each attribute is randomized in the tree. The advantages of this method can increase the accuracy value, resistant to outliers and efficient to store data. The feature selection process can work on large data with complex parameters effectively. Random Forest is used in decision tree-based classification and averaging the results, random forest techniques are randomly generated which are used to build the final model. This approach can be applied to different types of data and produces good models [13].

K-Nearest Neighbor

K-Nearest Neighbor (KNN) This algorithm is a method used for data classification based on the shortest distance of objects by determining the best K value [14]. A high K value reduces the effect of noise on classification and makes the boundaries more blurred. The K- Nearest Neighbor algorithm works by finding the closest K for pattern recognition using test data. Then KNN will calculate the probability of the test data belonging to class K and the class that has a high probability will be selected [15].

$$C = arg\ maxc\mathrm{i}\ (Sc(d0,\ C\mathrm{i}) \tag{1}$$

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a linear transformation to determine a new coordinate system from a dataset by reduction techniques or reducing large information data from images without removing the actual information [16] .The PCA algorithm has Eigenface as an image decomposition in a set of characteristic features, each image is implemented as a linear of eigen vectors. According to [17] in mathematical form it can be said that Y is a linear combination of variables X1,X2,...,Xp which can be interpreted as follows:

$$Y_1=W_{11}\, X_1 + W_{21}\, X_2 + W_{31}\, X_3 + \dots +W_{p1}\, X_p \qquad (2)$$
$$Y_2=W_{12}\, X_1 + W_{22}\, X_2 + W_{32}\, X_3 + \dots +W_{p2}\, X_{p.}$$
$$Yp=W_{pp}\, X_1 + W_{2p}\, X_2 + W_{3p}\, X_3 + \dots +W_{pp}\, X_p$$

Description:
Y = The estimated factor (a linear combination of variable x).
W= variable value weight or coefficient
P = sum of variables

Confusion Matrix

Confusion matrix is a method used in calculating accuracy in the concept of data mining, its function is to see the percentage of data mining accuracy, precision and recall [18]. The Confusion Matrix table has four combinations, namely *True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN)*. The performance evaluation of classification models includes Accuracy, Precision, Recall, Specificity and F1-Score [19]. The evaluation of binary classification distinguishes two approaches to binary attribute assignment, one of which is usually used as the standard approach, while the other is under investigation. The performance of classifiers or predictors can be evaluated using various metrics; some disciplines have for certain indicators due to various purposes [20]. [21] Find out whether the accuracy, sensitivity, and precision of a classification model's performance can be inferred from its parameters. The formulation of these performance metrics is as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \qquad (3)$$

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (5)$$

Receiver Operating Characteristic (ROC)

Classification model performance evaluation includes Accuracy, Precision, Recall, Specificity and F1-ScoreAUC value, also known as Receiver Operating Characteristic (ROC) value is a useful visual aid to compare two classification models. In ROC, the confusion matrix is expressed. The two-dimensional graph called ROC shows true positive results as vertical lines and false positive results as horizontal lines [22]. To graph or display computational results, the Confusion Matrix is represented as a ROC Curve. The region under the ROC line (Receiver Operating Characteristic) is known as AUC. Plotting the true positive rate against the false positive rate is done using a piecewise linear curve called ROC [23]. The trade-off between the model's ability to correctly identify positive tuples and its tendency to incorrectly identify negative tuples as positive tuples can be shown by the ROC curve [24][25]. When comparing the accuracy values of each categorization model graphically, the ROC curve works as well. The formula used is as follows:

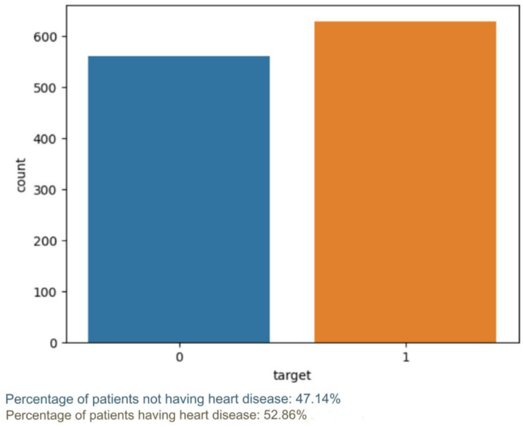
$$\text{TPR} = \frac{TP}{Actual\ Positive} = \frac{TP}{TP + FN} \qquad (6)$$

$$\text{FPR} = \frac{TP}{Actual\ Negative} = \frac{FP}{TN + FP} \qquad (7)$$

RESULTS AND DISCUSSIONS

Attributes that did not show a strong relationship with each other were eliminated to ascertain the relationship between each quality variable. Attribute distribution and data balance were examined using various data visualization techniques to improve understanding of the data set and its aspects. Afterwards, categorical data was converted to numerical for easy understanding and to simplify further data processing.

The next process, EDA, is used to identify missing values or duplicated data. Then visualization is performed to gain in-depth knowledge before proceeding with further data processing. The next process, EDA, is used to identify missing values or duplicated data. Then visualization is done to gain in-depth knowledge before proceeding with further data processing. The visualization data as shown in Figure 3.1 shows that 47.14% or 561 patients do not have heart disease, and 52.86% or 629 have heart disease.



Percentage of patients not having heart disease: 47.14%
Percentage of patients having heart disease: 52.86%

Figure 2. EDA distribution of patients having heart disease and not having heart disease

After performing EDA, researchers then performed PCA. The results of the principal component analysis, as illustrated by the Pareto plot, indicated that the optimal number of principal components was five, with an explained variance of 84%. The rationale behind this technique, which reduces the dimensionality of a dataset by focusing on the most important variables for analysis in multivariate data analysis, is based on the concept of cumulatively explained. The concept of cumulative explained variance is a fundamental aspect of PCA, which is a dimension reduction technique employed in multivariate data analysis. Subsequently, we utilized the dimension of the training and testing data for further use in modeling.

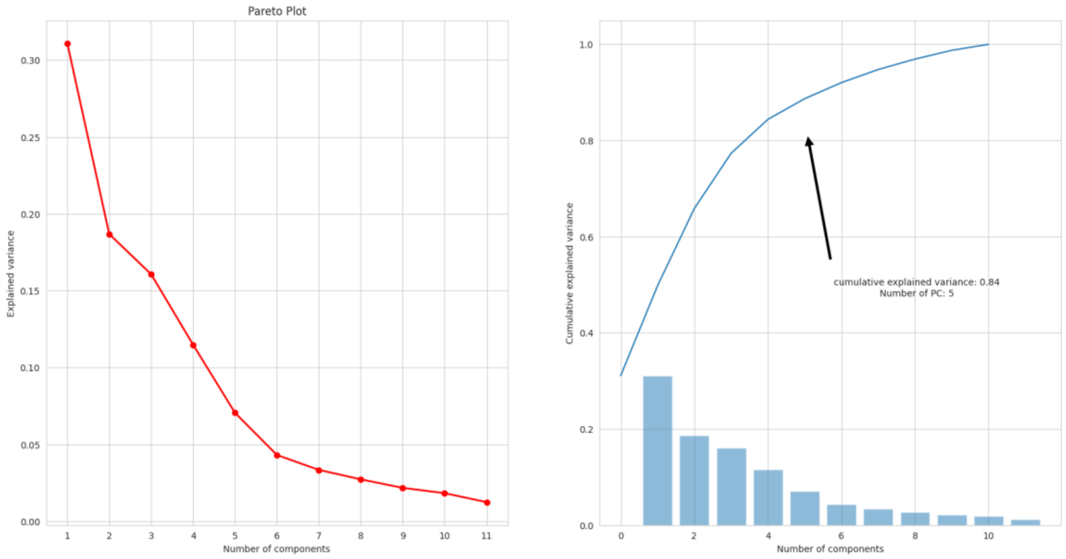


Figure 4. Pareto plot of eigenvalue in PCA

Then the algorithm modeling is carried out, with the results of the comparison of f-1, recall, accuracy, and precision values shown in Table 2**.** Comparing the results of Support Vector Machine, Random Forest, and K-Nearest Neighbor. The table shows that, the K-Nearest Neighbor method has better performance compared to Support Vector Machine and Random Forest.

Table 2. Result of modeling algorithm

Method	Accuracy	Precision	Recall	F1-Score
SVM	82%	79%	90%	82%
Random Forest	83%	82%	89%	86%
KNN	86%	85%	91%	88%

The table above shows the results of algorithm modeling, SVM algorithm with 82% accuracy, precision 0.79%, recall 0.90%, F1-Score 0.82%, while Random Forest has 83% accuracy, precision 0.82%, recall 0.89%, F1-Score 0.86%, and KNN algorithm with 86% accuracy, precision 0.85%, recall 0.91%, F1-Score 0.88%. Based on the algorithm modeling above, the best accuracy shows the KNN algorithm. The visualization confusion matrix of all models is depicted in Figure 5.

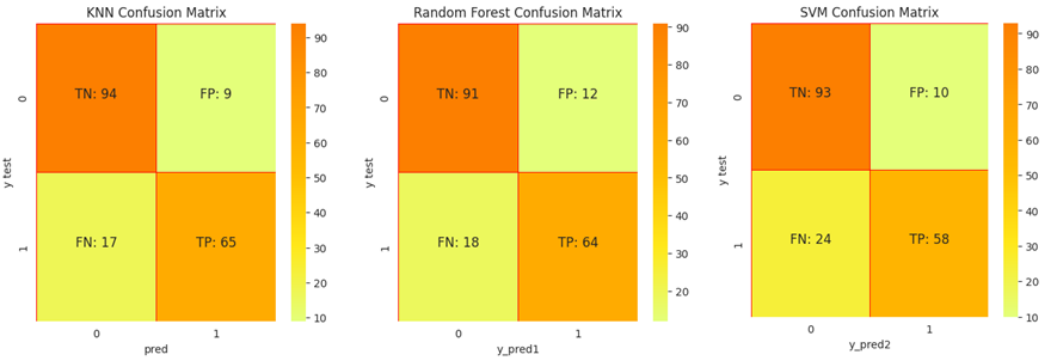


Figure 5. Confusion matrix of all model

The ROC curve can be generated by employing the True Positive Rate (TPR) and False Positive Rate (FPR) outcomes derived from confusion matrix calculations. The x-axis represents the TPR, and the t-axis represents the FPR. This study presents a comprehensive overview of false-positive and true-positive rates for all methods employed. It also presents the results of the false-positive and true-positive rates for all the methods used.

Table 3. Result of FPR and TPR each algorithm

Methode	FPR	TPR
SVM	0.0, 0.09708738,1.0	0.0,0.70731707,1.0
Random Forest	0.0, 0.11650485,1.0	0.0,0.7804878,1.0
KNN	0.0, 0.08737864,1.0	0.0,0.79268293,1.0

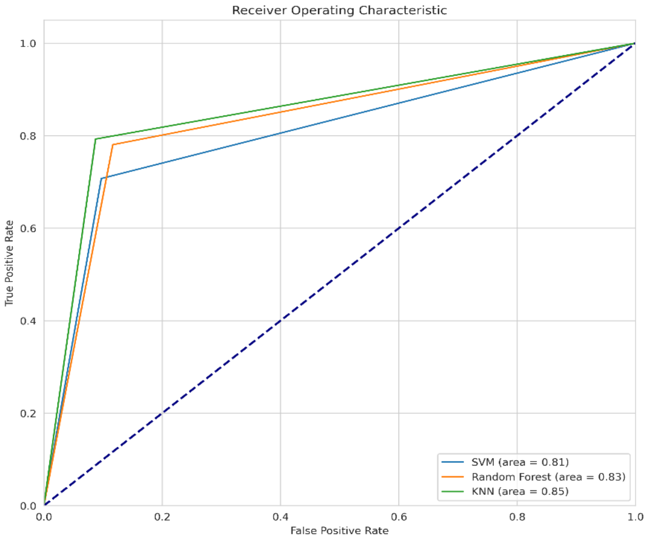


Figure 6. ROC of all model

With 3 methods used in the evaluation using ROC, the accuracy results are quite good. However, the best accuracy in this evaluation is K-Nearest Neighbor.

CONCLUSION

The results of research conducted through the preprocessing and PCA processes using testing and training data, as well as confusion matrix evaluation, indicate that the SVM modeling algorithm achieved an accuracy of 82%, the RF algorithm achieved an accuracy of 83%, and the KNN algorithm achieved an accuracy of 86%. These findings suggest that the KNN algorithm exhibited the highest accuracy. While the highest AUC results are found in the KNN algorithm with a result of 0.85, for the Random Forest algorithm 0.83 and the SVM algorithm 0.81. It can be concluded that the disease data classification algorithm. Based on the results of research that has been carried out through preprocessing and PCA processes using testing data and training data through confusion matrix evaluation, the resulting SVM modeling algorithm with 82% accuracy, RF with 83% accuracy, and KNN with 86% accuracy, it can be concluded that the best accuracy is the KNN algorithm. While the highest AUC results are found in the KNN algorithm with a result of 0.85, for the Random Forest algorithm 0.83 and the SVM algorithm 0.81. It can be concluded that the heart attack disease data classification algorithms that have been carried out by researchers produce quite good accuracy values. When evaluating various machine learning techniques, there is a big difference in the accuracy and AUC values. The accuracy rate looks good, but the best algorithm based on accuracy results is K-Nearest Neighbor which is the best algorithm among all algorithms. Future research is expected to implement it into an application system so that it is useful for the user community.

REFERENCES

[1] Z. Ahmadi, A. Abdullah, and I. Fakhruzi, “JURNAL MEDIA INFORMATIKA BUDIDARMA Meningkatkan Kemampuan Model dalam Memprediksi Penyakit Jantung dengan Algoritma NCL dan GridSearchCV,” vol. 7, pp. 1623–1633, 2023, doi: 10.30865/mib.v7i4.6142.

[2] L. Budianti and Suliadi, “Metode Weighted Random Forest dalam Klasifikasi Prediksi Kelangsungan Hidup Pasien Gagal Jantung,” *Bandung Conference Series: Statistics*, vol. 2, no. 2, pp. 103–110, Jul. 2022, doi: 10.29313/bcss.v2i2.3318.

[3] 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.

[4] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, “Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare,” *IEEE Access*, vol. 8, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

[5] A. A. Almazroi, E. A. Aldhahri, S. Bashir, and S. Ashfaq, “A Clinical Decision Support System for Heart Disease Prediction Using Deep Learning,” *IEEE Access*, vol. 11, pp. 61646–61659, 2023, doi: 10.1109/ACCESS.2023.3285247.

[6] A. Abdellatif, H. Abdellatef, J. Kanesan, C. O. Chow, J. H. Chuah, and H. M. Gheni, “An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyperparameter Optimization Methods,” *IEEE Access*, vol. 10, pp. 79974–79985, 2022, doi: 10.1109/ACCESS.2022.3191669.

[7] M. M. Ali *et al.*, “A machine learning approach for risk factors analysis and survival prediction of Heart Failure patients,” *Healthcare Analytics*, vol. 3, Nov. 2023, doi: 10.1016/j.health.2023.100182.

[8] Y. M. Ayano, F. Schwenker, B. D. Dufera, and T. G. Debelee, “Interpretable Machine Learning Techniques in ECG-Based Heart Disease Classification: A Systematic Review,” *Diagnostics*, vol. 13, no. 1, p. 111, Dec. 2022, doi: 10.3390/diagnostics13010111.

[9] K. M. Mohi Uddin, R. Ripa, N. Yeasmin, N. Biswas, and S. K. Dey, “Machine learning-based approach to the diagnosis of cardiovascular vascular disease using a combined dataset,” *Intell Based Med*, vol. 7, Jan. 2023, doi: 10.1016/j.ibmed.2023.100100.

[10] V. Chang, V. R. Bhavani, A. Q. Xu, and M. A. Hossain, “An artificial intelligence model for heart disease detection using machine learning algorithms,” *Healthcare Analytics*, vol. 2, Nov. 2022, doi: 10.1016/j.health.2022.100016.

[11] A. M. Qadri, A. Raza, K. Munir, and M. S. Almutairi, “Effective Feature Engineering Technique for Heart Disease Prediction With Machine Learning,” *IEEE Access*, vol. 11, pp. 56214–56224, 2023, doi: 10.1109/ACCESS.2023.3281484.

[12] M. S. Al Reshan, S. Amin, M. A. Zeb, A. Sulaiman, H. Alshahrani, and A. Shaikh, “A Robust Heart Disease Prediction System Using Hybrid Deep Neural Networks,” *IEEE Access*, vol. 11, pp. 121574–121591, Oct. 2023, doi: 10.1109/access.2023.3328909.

[13] Y. Pan, M. Fu, B. Cheng, X. Tao, and J. Guo, “Enhanced deep learning assisted convolutional neural network for heart disease prediction on the internet of medical things platform,” *IEEE Access*, vol. 8, pp. 189503–189512, 2020, doi: 10.1109/ACCESS.2020.3026214.

[14] N. Najwa Mohd Rizal, G. Hayder, M. Mnzool, B. M. E. Elnaim, A. O. Y. Mohammed, and M. M. Khayyat, “Comparison between Regression Models, Support Vector Machine (SVM), and Artificial Neural Network (ANN) in River Water Quality Prediction,” *Processes*, vol. 10, no. 8, Aug. 2022, doi: 10.3390/pr10081652.

[15] M. Azhari, Z. Situmorang, and R. Rosnelly, “Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes,” *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 2, p. 640, Apr. 2021, doi: 10.30865/mib.v5i2.2937.

[16] R. Supriyadi, W. Gata, N. Maulidah, A. Fauzi, I. Komputer, and S. Nusa Mandiri Jalan Margonda Raya No, “Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah,” vol. 13, no. 2, pp. 67–75, 2020, [Online]. Available: http://journal.stekom.ac.id/index.php/E-Bisnis■page67

[17] J. Homepage, S. R. Cholil, T. Handayani, R. Prathivi, and T. Ardianita, “IJCIT (Indonesian Journal on Computer and Information Technology) Implementasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Beasiswa,” 2021.

[18] S. R. Cholil, T. Handayani, R. Prathivi, and T. Ardianita, “Implementasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Beasiswa,” *IJCIT (Indonesian Journal on Computer and Information Technology)*, vol. 6, no. 2, Dec. 2021, doi: 10.31294/ijcit.v6i2.10438.

[19] D. Normawati and S. A. Prayogi, “IMPLEMENTASI NAÏVE BAYES CLASSIFIER DAN CONFUSION MATRIX PADA ANALISIS SENTIMEN BERBASIS TEKS PADA TWITTER,” *J-SAKTI (Jurnal Sains Komputer & Informatika)*, vol. 5, no. 2, 2021.

[20] A. M. Siregar, “Klasifikasi Untuk Prediksi Cuaca Menggunakan Esemble Learning,” *PETIR*, vol. 13, no. 2, pp. 138–147, Sep. 2020, doi: 10.33322/petir.v13i2.998.

[21] M. N. Winnarto, M. Mailasari, and A. Purnamawati, “KLASIFIKASI JENIS TUMOR OTAK MENGGUNAKAN ARSITEKTURE MOBILENET V2,” *Jurnal SIMETRIS*, vol. 13, no. 2, 2022.

[22] A. J. Bowers and X. Zhou, “Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes,” *J Educ Stud Placed Risk*, vol. 24, no. 1, pp. 20–46, Jan. 2019, doi: 10.1080/10824669.2018.1523734.

[23] S. Rampogu, “A review on the use of machine learning techniques in monkeypox disease prediction,” *Science in One Health*, vol. 2, p. 100040, 2023, doi: 10.1016/j.soh.2023.100040.

[24] A. U. Zailani and N. L. Hanun, “PENERAPAN ALGORITMA KLASIFIKASI RANDOM FOREST UNTUK PENENTUAN KELAYAKAN PEMBERIAN KREDIT DI KOPERASI MITRA SEJAHTERA,” *Infotech: Journal of Technology Information*, vol. 6, no. 1, pp. 7–14, Jun. 2020, doi: 10.37365/it.v6i1.61.

[25] H. Hozairi, A. Anwari, and S. Alim, “IMPLEMENTASI ORANGE DATA MINING UNTUK KLASIFIKASI KELULUSAN MAHASISWA DENGAN MODEL K-NEAREST NEIGHBOR, DECISION TREE SERTA NAIVE BAYES,” *Network Engineering Research Operation*, vol. 6, p. 133, Nov. 2021, doi: 10.21107/nero.v6i2.237.