Vol.7 No.1 Tahun 2026
Jurnal Kecerdasan Buatan, Komputasi dan Teknologi Informasi
https://ejournal.unuja.ac.id/index.php/core
E-ISSN : 2774-7875 and P-ISSN : 2775-0124

Application of the K-Nearest Neighbor (KNN) Algorithm in Data Mining for Heart Disease Prediction

Zaehol Fatah[1], Nur Aida[2]
[1,2] Universitas Ibrahimy Situbondo, Indonesia

Article Info

Article history:

Received : 12 03, 2025
Revised : 01 10, 2026
Accepted : 02 07, 2026

Keywords:

K-Nearest Neighbor;
Classification;
Heart disease;
Data mining;
RapidMiner;

Abstract

Cardiovascular (heart) disease is a leading cause of high global mortality rates. This is often exacerbated by low public awareness and limited access to cardiac screening facilities. To address these challenges, this study aims to apply the K-Nearest Neighbors (KNN) algorithm to classify heart disease and evaluate the resulting accuracy. The KNN algorithm was selected for its efficiency and ease of use in classifying large datasets; it works by measuring the distance between objects using the Euclidean distance formula. This classification model was built using RapidMiner 9.10 software. The data used was sourced from the Cleveland UCI Machine Learning Repository database available via Kaggle, consisting of a total of 303 patient records and characterized by 14 features used for prediction. The target variable (heart disease) was set as the classification label. Test results using cross-validation demonstrated that the KNN implementation is effective for predicting heart disease. This model achieved an accuracy of 64.03% with the parameter setting K=5. Further analysis of the confusion matrix identified 124 true-positive patients, with a maximum precision of 64.58%. Overall, these accuracy results confirm the potential and capability of the KNN method in classifying diagnostic data for heart disease patients.

Corresponding Author:

Nur Aida
ainuraaida807@gmail.com

INTRODUCTION

Cardiovascular (heart) disease is one of the leading causes of high mortality rates, alongside other serious conditions such as stroke, various types of cancer, and AIDS. Public awareness of the importance of heart health is relatively low. Limitations in medical services—such as a shortage of cardiologists and restrictive doctor schedules—often discourage people from getting their hearts checked. Medically speaking, heart disease is generally triggered by narrowing of the coronary arteries caused by atherosclerosis, spasm, or a combination of both conditions(Nugroho, 2021). Medically speaking, heart disease is generally caused by narrowing of the coronary arteries due to atherosclerosis, spasm, or a combination of both conditions(Nursahid et al., 2025).

Given these diagnostic challenges, an efficient method is needed. Data mining is defined as a series of steps aimed at exploring knowledge or information previously hidden within large volumes of data(Gautama & Arijanto, 2024). This discovery of new knowledge is achieved through the extraction and selection of data attributes that are relevant and influential to predictive capabilities. This process utilizes various techniques, including statistics, artificial intelligence (AI), and machine learning.(Khodijah et al., 2024)

The field of data mining offers a variety of algorithms that play a crucial role in data analysis and the extraction of new, useful knowledge(Permatasari et al., 2024). One of the most widely used and popular algorithms among researchers is the K-Nearest Neighbors (KNN) algorithm. KNN is known for its simplicity yet effectiveness with large datasets; its function is to classify new data based on proximity to the nearest objects (neighbors)(Alici Davutoglu, 2023). Determining the number of nearest neighbors (K) significantly influences classification decisions in the KNN method. Since it falls under supervised learning, the dataset used must include a target variable (label)(Siregar & Wijayanti, 2025). In the context of this research, the value of K is calculated using the Euclidean Distance formula(Putra & Supriyanto, 2023). Therefore, this study focuses primarily on applying the KNN method to classify heart disease while evaluating the level of accuracy that can be achieved by this method.

METHOD

Data mining is a procedure for extracting and identifying valuable information by applying disciplines such as mathematics, statistics, artificial intelligence, and machine learning. In short, data mining is the activity of discovering hidden patterns in a dataset. Based on its primary functions, data mining is divided into several categories: description, estimation, prediction, classification, clustering, and association(Gautama & Arijanto, 2024). The data mining process generally involves three essential stages. The crucial initial step is data selection, cleaning, and preprocessing. This stage is carried out based on specific guidelines and expertise from the relevant domain, encompassing the capture and integration of data (both internal and external) to create a comprehensive picture of the organization. Once the data is integrated, data mining algorithms are then applied to explore the dataset, facilitating the identification of valuable information. However, this data processing often requires a significant amount of time(Rahmadini et al., 2023).

Machine learning is defined as a collection of algorithms used to optimize the performance of a system or computer based on sample data. The primary strength of machine learning lies in its ability to modify and adjust decision-making to adapt to changing conditions(S. C. Dewi et al., 2024). The main advantage of machine learning is its flexibility in processing and adapting data to incoming changes. In the context of applications, machine learning has several key uses(Islam et al., 2023):

a. Classification is a technique aimed at predicting values or determining the class of each individual within a population(Tampubolon, 2023).
b. Similarity matching is a procedure used in machine learning to identify the degree of similarity among individuals based on available data(Low et al., 2023).
c. Clustering is a machine learning method that divides data into several groups based on specific criteria or characteristics(Wadhwa, 2021).

K-Nearest Neighbor (K-NN) is an algorithm used to classify a dataset based on the principle of learning from previously classified data. This algorithm falls under supervised learning, in which a new instance (query) is classified into a category based on the majority of the closest distances to existing data points(Adi & Wintarti, 2022). The KNN process involves finding a set of K objects in the training data that have the highest similarity or are closest in distance to the object being tested (new data). To carry out this process effectively, the classification system must be capable of efficiently retrieving information(Nainggolan et al., 2025). In the KNN method, geometric distance formulas are used to calculate proximity values. One commonly used formula is the Euclidean distance formula, which is known for its simplicity and its ability to produce high classification accuracy(Hakim et al., 2025):

DOI: https://doi.org/10.33650/coreai.v6i2

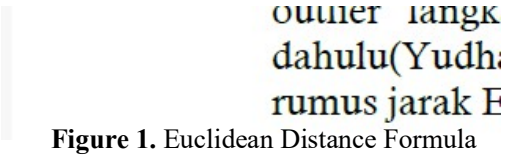

Figure 1. Euclidean Distance Formula

Classification is a data mining method used to group data into predefined categories or classes based on patterns and attributes found in the dataset(Rahmadan & Shudiq, 2024). This technique is becoming increasingly popular because it can process a broader range of data compared to regression methods(Nasien et al., 2024). The main objective of classification is to predict the category (class or label) of unidentified data based on information and patterns obtained from previously known training data(Homaidi & Fatah, 2024).

The data collection process for this study utilized publicly available heart health data on the kaggle.com platform. The dataset can be accessed via the link *https://www.kaggle.com/code/dirgahalimsusilo/klasifikasi-penyakit-jantung*. The dataset used consists of raw data sourced from the Cleveland database in the UCI Machine Learning Repository. A total of 303 patient records were used in this dataset. Each record is characterized by 14 attributes or features that play a key role in predicting the patients' condition—whether or not they have heart disease. Details of the dataset and a description of each attribute are presented in Table 1(Andiani et al., 2020).

Table 1. Dataset

No.	*age*	*sex*	*chol*	*cp*	*tresbps*	*fbs*	*restejg*	*thalach*	*exang*	*oldpeak*	*slope*	*ca*	*thal*	*target*
1	63	1	233	3	145	1	0	150	0	23	0	0	1	1
2	37	1	250	2	130	0	1	187	0	35	0	0	2	1
3	41	0	204	1	130	0	0	172	0	14	2	0	2	1
4	56	1	236	1	120	0	1	178	0	8	2	0	2	1
5	57	0	354	0	120	0	1	163	1	6	2	0	2	1
6	57	1	192	0	140	0	1	148	0	4	1	0	1	1
7	56	0	294	1	140	0	0	153	0	13	1	0	2	1
8	44	1	263	1	120	0	1	173	0	0	2	0	3	1
9	52	1	199	2	172	1	1	162	0	5	2	0	3	1
10	57	1	168	2	150	0	1	174	0	16	2	0	2	1
11	54	1	239	0	140	0	1	160	0	12	2	0	2	1
12	48	0	275	2	130	0	1	139	0	2	2	0	2	1
13	49	1	266	1	130	0	1	171	0	6	2	0	2	1
14	64	1	211	3	110	0	0	144	1	18	1	0	2	1
15	58	0	283	3	150	1	0	162	0	1	2	0	2	1
290	55	0	205	0	128	0	2	130	1	2	1	1	3	0
291	61	1	203	0	148	0	1	161	0	0	2	1	3	0
292	58	1	318	0	114	0	2	140	0	44	0	3	1	0
293	58	0	225	0	170	1	0	146	1	28	1	2	1	0
294	67	1	212	2	152	0	0	150	0	8	1	0	3	0
295	44	1	169	0	120	0	1	144	1	28	0	0	1	0
296	63	1	187	0	140	0	0	144	1	4	2	2	3	0
297	63	0	197	0	124	0	1	136	1	0	1	0	2	0
298	59	1	176	0	164	1	0	90	0	1	1	2	1	0
299	57	0	241	0	140	0	1	123	1	2	1	0	3	0
300	45	1	264	3	110	0	1	132	0	12	1	0	3	0
301	68	1	193	0	144	1	1	141	0	34	1	2	3	0
302	57	1	131	0	130	0	1	115	1	12	1	1	3	0
303	57	0	236	1	130	0	0	174	0	0	1	1	2	0

Description:

Age : patient's age.
Sex : gender (1 = male, 0 = female)
Chol : serum cholesterol.
Cp : chest pain (0 = chest pain reducing blood supply to the heart, 1 = chest pain not related to the heart, 2 = esophageal spasm (not related to the heart), 3 = chest pain showing no signs of disease)
Restecg : resting electrocardiogram results (0 = normal, 1 = mild to severe issues)
Fbs : fasting blood sugar (1 = if FBS is greater than 120 mg/dL, 0 = if not)

Trestbps : blood pressure (blood pressure above 130–140 falls into the concerning category)
Slope : slope of the peak exercise ST segment (0 = upsloping, 1 = flatsloping, 2 = downsloping)
Exang : patient's condition during physical activity or exercise (0 = no pain, 1 = causes pain)
Oldpeak : relative exercise-induced ST depression
Ca : number of major blood vessels
Thalach : patient's maximum heart rate
Thal : thallium stress test
Target : presence or absence of disease (1 = yes, 0 = no)

This research focuses on designing a machine learning model to classify patients' cardiovascular (heart) disease history. K-Nearest Neighbors (KNN) was selected as the classification method and implemented using RapidMiner version 9.10. This study is supported by a dataset obtained from the Kaggle platform, originally sourced from the Cleveland database at the UCI Machine Learning Repository. The dataset contains data on 303 patients and includes 14 attributes/features used as classification parameters, with the "cardiov" (heart disease) label serving as the indicator variable for the presence of heart disease.

RESULTS AND ANALYSIS

This research focuses on designing a machine learning model to classify patients' cardiovascular (heart) disease history. The K-Nearest Neighbors (KNN) method was selected for classification and implemented using RapidMiner version 9.10. This study is supported by a dataset obtained from the Kaggle platform, originally sourced from the Cleveland database at the UCI Machine Learning Repository. The dataset contains data on 303 patients and includes 14 attributes/features used as classification parameters, with the "cardiov" (heart disease) label serving as the indicator variable for the presence of heart disease.

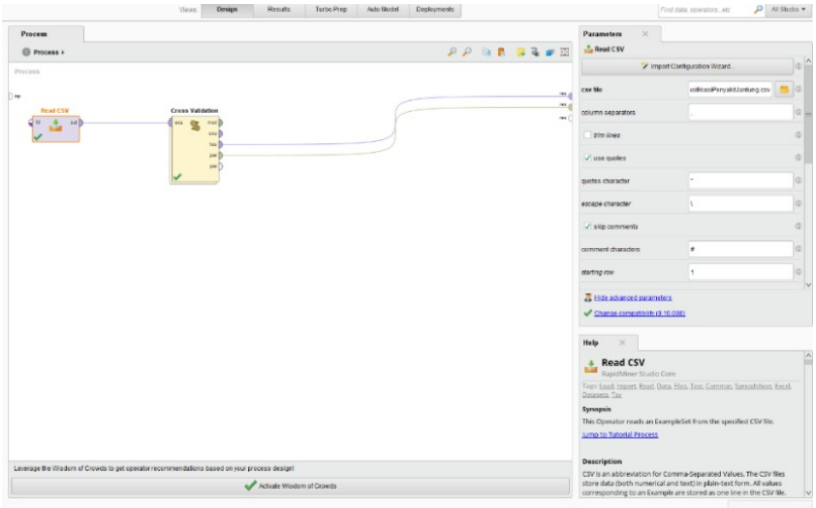

Figure 2.Data Transformation in RapidMiner

This data transformation is a crucial step for initializing the values in the dataset and ensuring that the data types meet the requirements of the K-Nearest Neighbors (KNN) method. Through this process, one of the criteria or attributes must also be designated as the label variable (dependent variable)(Afdal & Elita, 2022). In the dataset used in this study, the target variable was designated to serve as the classification label.

DOI: https://doi.org/10.33650/coreai.v6i2

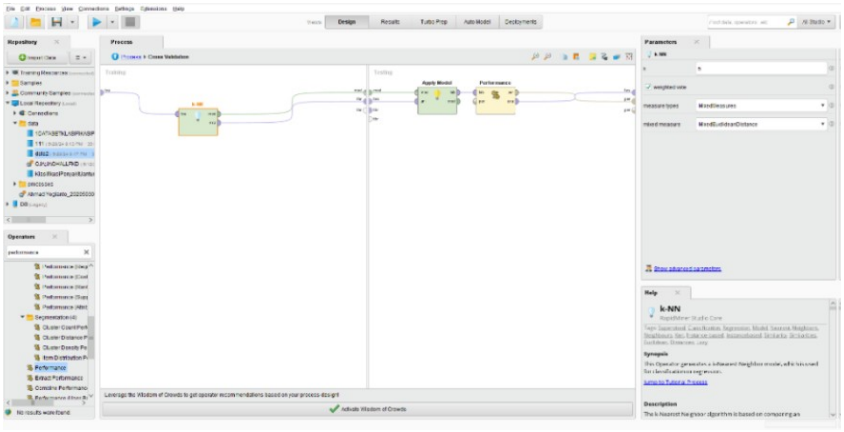

Figure 2. Model method K-Nearest Neighbor (KNN)

The data mining process in this study was implemented using the RapidMiner application with the K-Nearest Neighbors (KNN) algorithm. The designed process involves a series of operators, as illustrated in Figure 3. The initial stage begins with the "Read CSV" operator, which reads and imports the dataset stored in CSV format into the process flow. Next, to evaluate the model's objectivity and performance, the "Cross Validation" operator is used(Arif, 2024).

Cross-validation is an essential statistical technique for testing the performance of the method used. In practice, this process divides the dataset into two main subsets: training data and testing data(S. P. Dewi et al., 2022). The testing dataset subset also includes the performance operator. This operator plays a crucial role in evaluating the model's performance. The performance operator provides a complete list of performance metrics (such as accuracy, precision, and others) for the model being tested.

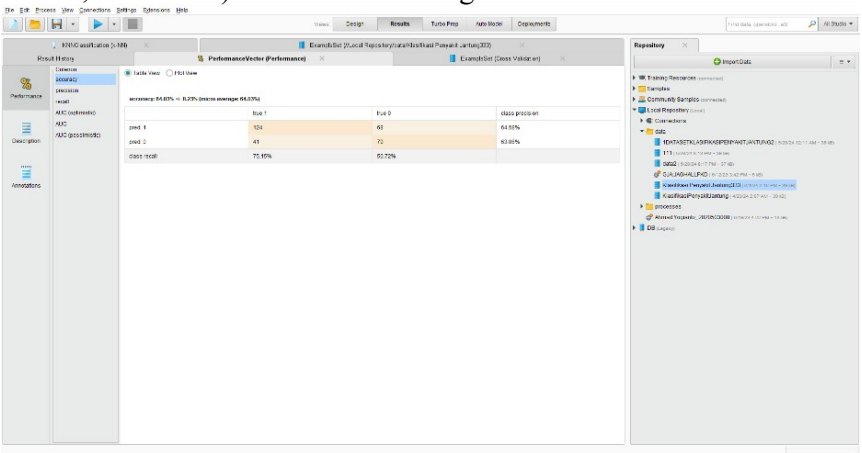

Figure 2. Proces K-Nearest Neighbor (KNN)

Based on the results presented, the implementation of the K-Nearest Neighbor (KNN) method with a parameter setting of k = 5 showed a model accuracy of 64.03%. Further analysis of the confusion matrix shows that 124 patients were predicted to have heart disease and tested positive (true positives). Conversely, 68 patients were predicted to have heart disease but tested negative (false positives). Meanwhile, there were 70 patients who were predicted not to have heart disease, and the results were negative (true negatives). From the matrix evaluation, the highest precision value achieved was 64.58%. The recall metric showed a result of 75.15% for patients with heart disease and 50.72% for patients without heart disease.

CONCLUSIONS

The application of the K-Nearest Neighbor (KNN) algorithm for classifying heart disease in the Cleveland dataset from the UCI Machine Learning Repository using RapidMiner version 9.10 has been successfully implemented, thereby fulfilling the research objective. Test results show that the KNN method is effective for predicting heart disease, with an accuracy rate of 64.03% at a K value of 5. This accuracy indicates that the KNN method is capable of classifying patient data into those with and without heart disease, although there is still room for performance improvement.

Further research should explore varying values of K to optimize the accuracy of the classification model. In addition, to further refine the research that has been conducted, it is also recommended to conduct a comparative study between the K-Nearest Neighbor algorithm and other machine learning algorithms, such as the support vector machine (SVM) or Naive Bayes, in order to determine which method is the most effective and provides the best predictive results for heart disease.

REFERENCES

Adi, S., & Wintarti, A. (2022). Komparasi metode support vector machine (SVM), K-Nearest Neighbors (KNN), Dan Random Forest (RF) untuk prediksi penyakit gagal jantung. *MATHunesa: Jurnal Ilmiah Matematika*, *10*(2), 258–268. https://doi.org/10.26740/mathunesa.v10n2.p258-268

Afdal, M., & Elita, L. R. (2022). Penerapan Text Mining Pada Aplikasi Tokopedia Menggunakan Algoritma K-Nearest Neighbor. *Jurnal Ilmiah Rekayasa Dan Manajemen Sistem Informasi*, *8*(1), 78–87. https://doi.org/10.30865/jurikom.v13i3

Alici Davutoglu, E. (2023). MEDICAL BIOCHEMISTRY. INTERNATIONAL JOURNAL OF MEDICAL BIOCHEMISTRY, 6(1), 1-10. https://doi.org/10.30865/mib.v8i3.7844

Andiani, L., Sukemi, S., & Rini, D. P. (2020). Analisis penyakit jantung menggunakan metode knn dan random forest. Annual Research Seminar (ARS), 5(1), 165-169. https://doi.org/10.32639/p5e7b161

Arif, S. N. N. (2024). Klasifikasi Penyakit Serangan Jantung Menggunakan Metode Machine Learning K-Nearest Neighbors (KNN) dan Support Vector Machine (SVM). Jurnal Media Informatika Budidarma, 8(3). https://doi.org/10.30865/mib.v10i1

Dewi, S. C., Putra, C. E., & Nugraheni, A. G. (2024). Implementasi Metode K-Nearest Neighbors (KNN) dan Naive Bayes untuk Klasifikasi Penyakit Jantung. Technology and Informatics Insight Journal, 3(2), 76-94. https://doi.org/10.32639/p5e7b161

Dewi, S. P., Nurwati, N., & Rahayu, E. (2022). Penerapan Data Mining Untuk Prediksi Penjualan Produk Terlaris Menggunakan Metode K-Nearest Neighbor. Build. Informatics, Technol. Sci, 3(4), 639-648. https://doi.org/10.47065/bits.v3i4.1408

Gautama, K., & Arijanto, R. (2024). PENERAPAN DATA MINING UNTUK PREDIKSI PENJUALAN PRODUK TERLARIS MENGGUNAKAN METODE K-NEAREST NEIGHBOR PADA CV. NUANSA KARUNA LESTARI. JATI (Jurnal Mahasiswa Teknik Informatika), 8(6), 11184-11190. https://doi.org/10.36040/jati.v8i6.11189

Hakim, L., Sobri, A., Sunardi, L., & Nurdiansyah, D. (2025). Prediksi penyakit jantung berbasis mesin learning dengan menggunakan metode k-nn. Jurnal Digital Teknologi Informasi, 7(2), 14-20. https://doi.org/10.32502/digital.v7i2.9429

Homaidi, A., & Fatah, Z. (2024). Implementasi Metode K-Nearest Neighbors (KNN) untuk Klasifikasi Penyakit Jantung. G-Tech: Jurnal Teknologi Terapan, 8(3), 1720-1728. https://doi.org/10.33379/gtech.v8i3.4495

Islam, A., Mahmood, Z., & Khan, U. (2023). Double-diffusive stagnation point flow over a vertical surface with thermal radiation: Assisting and opposing flows. Science Progress, 106(1), 00368504221149798. https://doi.org/10.1177/00368504221149798

Khodijah, K., Sriyanto, S., Aziz, R. Z. A., & Suhendro, S. (2024). Perbandingan Kinerja Lima Algoritma Klàsifikasi Dasar Untuk Prediksi Penyakit Jantung "CLASSIFIER: NB, DTC4. 5, KNN, ANN & SVM". Jaringan Sistem Informasi Robotik-JSR, 8(2), 230-234. https://doi.org/10.31539/intecoms.v7i5.12434

Low, I., Shu, J., Xiao, M.-L., & Zheng, Y.-H. (2023). Amplitude/operator basis in chiral perturbation

theory. Journal of High Energy Physics, 2023(1), 1-33. https://doi.org/10.1007/JHEP01(2023)031

Nainggolan, J. K., Sinaga, F., Sitorus, A. M., Khairia, A., & Wijaya, B. A. (2025). Analisa Komparasi dengan Algoritma K-Nearest Neighbor (KNN) dan Support Vector Machine (SVM) untuk Prediksi Penyakit Jantung. Dinamik, 30(2), 297-306. https://doi.org/10.35315/dinamik.v30i2.10254

Nasien, D., Sirvan, S., Deny, D., Syahputra, R. S. R., Marunduri, A. A., & See, R. P. (2024). Klasifikasi Penyakit Jantung Menggunakan Decision Tree dan KNN Menggunakan Ektraksi Fitur PCA. JEKIN-Jurnal Teknik Informatika, 4(1), 18-24. https://doi.org/10.58794/jekin.v4i1.641

Nugroho, F. A. (2021). Perancangan sistem pakar diagnosa penyakit jantung dengan metode forward chaining. Jurnal Informatika Universitas Pamulang, 3(2), 75-79. https://doi.org/10.32493/informatika.v3i2.1431

Nursahid, W., Nugroho, B. I., & Syefudin, S. (2025). Optimalisasi Preprocessing Data Menggunakan Pendekatan CRISP-DM untuk Meningkatkan Kualitas Klasifikasi Penyakit Jantung. RIGGS: Journal of Artificial Intelligence and Digital Business, 4(3), 3621-3626. https://doi.org/10.31004/riggs.v4i3.2514

Permatasari, U. O. R., Jasri, M., & Shu. (2024). Prediksi Kelayakan Mahasiswa sebagai Penerima Beasiswa Bank Indonesia pada Tahap Seleksi Administrasi di Universitas Nurul Jadid menggunakan Algoritma K Nearest Neighbor. Journal of Electrical Engineering and Computer (JEECOM), 6(1), 252-260. https://doi.org/10.33650/jeecom.v6i1.8425

Putra, E. A., & Supriyanto, R. (2023). Analisis Algoritma Decesion Tree, KNN dan Naïve Bayes Pada Dataset Penyakit Jantung. Jurnal Ilmiah KOMPUTASI, 22(2), 273-284. https://doi.org/10.32409/jikstik.22.2.3353

Rahmadan, I., & Shudiq, W. J. (2024). Increasing student interest in learning through the implementation of the K-nearest neighbor algorithm in classifying learning preferences at SMAN 1 Kraksaan. Journal of Computer Networks, Architecture and High Performance Computing, 6(4), 1851-1862. https://doi.org/10.47709/cnahpc.v6i4.4526

Rahmadini, R., LorencisLubis, E. E., Priansyah, A., RWN, Y., & Meutia, T. (2023). Penerapan data mining untuk memprediksi harga bahan pangan di Indonesia menggunakan algoritma K-Nearest Neighbor. Jurnal Mahasiswa Akuntansi Samudra, 4(4), 223-235. https://doi.org/10.33059/jmas.v4i4.7074

Siregar, J., & Wijayanti, D. (2025). Integrasi Model Neural Network dengan SVM dan KNN untuk Deteksi Dini Penyakita Jantung pada Penderita Hypertensi. Technologia: Jurnal Ilmiah, 16(1), 104-109. https://doi.org/10.31602/tji.v16i1.17046

Tampubolon, G. C. (2023). ASSOCIATION BETWEEN SERUM ALP LEVELS AND PRE-OPERATIVE LDH TOWARDS CLINICAL PROGNOSIS OF OSTEOSARCOMA PATIENTS IN H. ADAM MALIK HOSPITAL, MEDAN. Sumatera Medical Journal, 6(1). https://doi.org/10.32734/sumej.v6i1.8991

Wadhwa, S. (2021). Performance comparison of classifiers on twitter sentimental analysis. Proceedings of The 5th International Conference on Applied Research in Science, Technology and Knowledge, 50-61. https://doi.org/10.33422/3rd.ictle.2021.02.134