KHATULISTIWA SMART : JOURNAL OF ARTIFICIAL INTELLIGENCE
e-ISSN : 3089-9710
p-ISSN : XXXX-XXXX

Khatulistiwa Smart : Journal of Artificial Intelligence
Vol. 2 No. 1 (2025) : 97 - 102
Available online at : https://journal.literasikhatulistiwa.org/index.php/smart

Hybrid Method Optimization For Classifying Heart Disease Using Knn And Pca Algorithms Based On Web Streamlite

Khofiyatul Hasanah [1] , Moh. Ainol Yaqin [2] , Cahyuni Novia [3]
[123] Universitas Nurul Jadid
khofiyatul27@gmail.com[1], ainolyaqin.unuja.ac.id[2], cahyuninovia@gmail.com[3]

Doi
Submitted: 26 Oct 2025 | Revision: 29 Nov 2025 | Accepted: 15 Dec 2025 | Published: 31 Dec 2025

Abstract Heart disease is one of the leading causes of death worldwide and often goes undetected early. This necessitates a decision support system capable of facilitating rapid and accurate diagnosis. This study aims to develop a heart disease classification system by combining two methods: K-Nearest Neighbors (KNN) and Principal Component Analysis (PCA), in a web-based application using Streamlit. PCA is used to reduce data dimensionality and eliminate less relevant features to improve classification efficiency and performance. Meanwhile, the KNN algorithm is used to determine the class (heart disease or not) based on the proximity of the new data to the labeled data. This study used the heart.csv dataset and was tested using several methods, including accuracy, classification reports, and confusion matrices. The test results showed that the hybrid PCA and KNN model was capable of providing relatively high accuracy and informative visualizations. The best accuracy rate achieved in this study reached 90%, demonstrating the model's effectiveness in classifying data. Using the Streamlit interface, this system is easily accessible and usable by users without requiring special installation. The conclusion of this study is that the combination of PCA and KNN is effective in classifying heart disease efficiently and accurately.

Keywords *Heart disease, KNN, PCA, classification, Streamlit*.

1. INTRODUCTION

Coronary heart disease (CHD) is one of the most common cardiovascular diseases and a leading cause of death worldwide (Saptadi et al., 2023). Premature deaths due to heart disease account for approximately 4% in high-income countries and 42% in low-income countries. It is estimated that by 2030, the number of deaths from heart disease will continue to increase, reaching 23.3 million people (Indonesian Ministry of Health, 2014). The main factors contributing to heart disease can be caused by many unhealthy habits, such as lack of physical activity, a diet high in salt and fat, smoking, excessive alcohol consumption, and

stress. Diabetes, obesity, high blood pressure, high cholesterol, and a family history of heart disease are additional risk factors that contribute to heart disease (Naomi et al., 2021).

This study aims to apply data mining techniques to improve the effectiveness of analysis and identify factors contributing to heart disease using Principal Component Analysis (PCA). PCA is used to identify and reduce the dimensions of heart disease factors. Using the PCA method, two main factors were identified that explain 75% of the total variance in the data. This study found that the PCA method is effective in reducing data dimensions and identifying the main factors underlying the data (Manullang et al., 2024).

K-Nearest Neighbor (KNN) is one of the methods used to classify new objects by considering a number of K nearest neighbors. The KNN algorithm is relatively simple and easy to understand, so it is often used in various applications. In this algorithm, the classification of an image is based on the closest distance to its neighbors. This distance value will be used as a similarity value between the test data and the training data (Akbarollah et al., 2023). Hasran's research results using the KNN algorithm on a heart disease dataset obtained from the UCI Machine Learning Repository dataset center obtained a K=6 value with an accuracy of 85%, a precision value of 78%, a recall of 93%, and an f-measure of 85% (Hasran, 2020). The K-Nearest Neighbor (KNN) algorithm is a method for finding groups of training data objects with the highest similarity to test data objects (Lestari, 2014).

In addition to the classification process, other approaches are also needed to identify heart disease, including the Principal Component Analysis (PCA) method. PCA is used to recognize certain patterns in data and to highlight differences or similarities found in a set of data. Generally, this method is used as a tool to reduce data dimensions, transforming them into another form of representation in a different value space (Raysyah et al., 2021). PCA is an orthogonal linear transformation that serves to convert data into a new coordinate system, where the first coordinate, known as the first principal component, has the highest variability related to the projection, while the second coordinate has the second highest variability, and so on. The PCA process involves determining the standard deviation, covariance matrix, eigenvalues, and eigenvectors. PCA can be performed using the covariance or correlation method (Hasym & Susilawati, 2021). PCA is capable of reducing data complexity without losing important information, thereby significantly strengthening KNN performance in classification tasks (Zheng et al., 2023). The implementation of PCA before KNN has been shown to improve computational efficiency and prediction accuracy, especially in the medical and pattern recognition domains (Farhan & Hasan, 2023).

Based on the dataset obtained from the Kaggle website entitled Heart Disease Dataset written by M Yasser H, there are 14 features used in diagnosing heart disease, namely age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, target. The number of features is quite large, so a heart disease classification system is needed to produce more effective and accurate diagnoses. I decided to use the Python application.

Selecting the appropriate method for classifying heart disease is essential because it affects the results that will be displayed. Among the many algorithms that are frequently used and popular among researchers, we decided to use two methods, namely the K-Nearest Neighbors (KNN) method, the results of which will be compared with the results calculated using the Principal Component Analysis method. This study aims to implement the K-Nearest Neighbors (KNN) and Principal Component Analysis methods in heart disease classification and determine the accuracy level of these methods, as well as identify which method is the most accurate. This research is expected to improve the accuracy of early detection of heart disease, thereby helping medical practitioners make faster and more

accurate decisions. In addition, the results of this study are expected to provide new insights for the development of data-based diagnostic methods (Dewi et al., 2024).

Although previous studies have applied KNN and PCA algorithms in heart disease classification, most of them were conducted separately, resulting in less than optimal outcomes. KNN has weaknesses when dealing with high-dimensional data because it can reduce prediction accuracy, while PCA, although effective in reducing dimensions, is generally only used as a preprocessing stage without being thoroughly evaluated for its impact on classification performance. Furthermore, most studies are still limited to dataset-based trials without implementation in the form of applications that are easily accessible to non-technical users. Therefore, this study offers a new approach by combining the KNN and PCA methods in a hybrid manner and implementing them in an interactive web platform based on Streamlit. The uniqueness of this study lies in its attempt to compare the performance of KNN with and without PCA, while also presenting a web-based system that can help medical personnel detect heart disease early in a faster, more accurate, and more practical manner.

2. RESEARCH METHODS

This study uses a quantitative computational approach by analyzing heart disease datasets through machine learning algorithms. This research is experimental in nature, where models are developed, trained, and tested using secondary datasets from Kaggle. The data undergoes preprocessing, including cleaning, normalization, and label coding, before being used for model training. The main algorithms applied were K-Nearest Neighbor (KNN) and Principal Component Analysis (PCA) because both are effective in processing high-dimensional data.

The next step is to develop a web-based classification system using Streamlit with Python. Model performance is evaluated using a Confusion Matrix through four main metrics: accuracy, precision, recall, and F1-score. These evaluation results are used as a reference to assess the accuracy and reliability of the system in supporting heart disease detection.

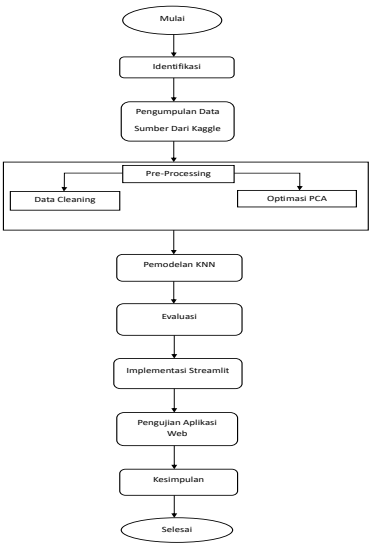

Figure 1. Research Flow Diagram

2.1. Dataset

The dataset used in this study was obtained from Kaggle under the title "Heart Disease Dataset," containing a total of 304 data points. The dataset source can be accessed via the website https://www.kaggle.com/datasets/yasserh/heart-disease-dataset/code.

2.2. Data Preprocessing

a. Data Cleaning

Data cleaning is a step in information processing that aims to correct, delete, or remove inaccurate, damaged, irrelevant, duplicate, or incomplete data from a data set. The purpose of data cleaning is to improve data quality so that it can be used effectively in data analysis.

b. PCA Optimization

PCA (Principal Component Analysis) optimization refers to the process of improving or refining the application of PCA to produce the most effective representation of data in accordance with the objectives of the analysis. In general, PCA is a dimension reduction technique used to convert high-dimensional data into lower-dimensional data, while retaining as much information (variation) as possible in the data.

2.3. Split Dataset

The variables included in this sample data cover the type of chest pain, resting blood pressure, cholesterol level, fasting blood sugar level, resting electrocardiogram results, maximum heart rate, and chest pain (angina) triggered by physical activity. In addition, there are also depression levels, ST segment depression, number of colored blood vessels, thalassemia, and the presence of heart disease.

3. RESULTS

The variables included in this sample data include chest pain type, resting blood pressure, cholesterol level, fasting blood sugar level, resting electrocardiogram results, maximum heart rate, and chest pain (angina) triggered by physical activity. In addition, there are also depression levels, ST segment inclination, number of colored blood vessels, thalassemia, and the presence of heart disease. The K-Nearest Neighbor (KNN) and Principal Component Analysis (PCA) methods are applied to support the heart disease classification process, which will be integrated into a web-based platform using Streamlit.

Table 1. Number of Datasets

No	Heart Disease	Total
1	Age	16473
2	sex	207
3	cp	293
4	trestbps	39882
5	chol	74618
6	fbs	45
7	restecg	160
8	thalach	45343
9	exang	99
10	oldpeak	315
11	slope	424

12	ca	221
13	thal	701
14	target	165
Total		**162473**

4. CONCLUSIONS

K-Nearest Neighbors (KNN) modeling with PCA dimension reduction can improve classification efficiency and accuracy. Euclidean distance is used to determine the proximity between data, while evaluation through a confusion matrix shows the model's performance in terms of accuracy, precision, recall, and F1-score. Implementation into a Streamlit-based application makes it easy for users to perform interactive predictions, and web testing ensures that the system runs stably and is suitable for use.

5. REFERENCES

References cited during the preparation of the article should be included in the bibliography. For efficiency and brevity purposes, the number of references used should be no more than 40 and no less than 10, with a proportion of 70% journal references, preferably international journals, and 30% book references. Journal references should be from the last 5 years, while book references can be more flexible. This journal uses the American Psychological Association 7th format.

Saptadi, J. D., Arianto, M. E., & Labibah, L. (2023). Edukasi Pencegahan Penyakit Jantung Melalui Media Poster Di Rt 05 Dusun Gebang, Sleman, Diy. *IJECS: Indonesian Journal of Empowerment and Community Services*, *4*(1), 30–34.Kilaru, R., & Raju, K. M. (2022). Prediction of Maize Leaf Disease Detection to improve Crop Yield using Machine Learning based Models. *4th International Conference on Recent Trends in Computer Science and Technology, ICRTCST 2021 - Proceedings*, 212–217. https://doi.org/10.1109/ICRTCST54752.2022.9782023

Naomi, W. S., Picauly, I., & Toy, S. M. (2021). Faktor Risiko Kejadian Penyakit Jantung Koroner. *Media Kesehatan Masyarakat*, *3*(1), 99–107. https://doi.org/10.35508/mkm.v3i1.3622

Manullang, S., Kairani, N., Sinaga, M. S., & Hutapea, B. (2024). *ANALISIS FAKTOR PENYEBAB PENYAKIT JANTUNG MENGGUNAKAN METODE PRINCIPAL COMPONENT ANALYSIS (PCA) masyarakat . Penyakit jantung juga merupakan masalah kesehatan yang kritis karena terikat , melainkan mencari saling ketergantungan antar variabel untuk meng*. *5*(3).

Raysyah, S. R., Veri Arinal, & Dadang Iskandar Mulyana. (2021). Klasifikasi Tingkat Kematangan Buah Kopi Berdasarkan Deteksi Warna Menggunakan Metode Knn Dan Pca. *JSiI (Jurnal Sistem Informasi)*, *8*(2), 88–95. https://doi.org/10.30656/jsii.v8i2.3638

Hasym, I. E., & Susilawati, I. (2021). Konvergensi Teknologi dan Sistem Informasi Klasifikasi Jenis Ikan Cupang Menggunakan Algoritma Principal Component Analysis (PCA) Dan K-Nearest Neighbors (KNN). *KONSTELASI: Konvergensi Teknologi Dan Sistem Informasi*, 168–179.

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