International Journal Science and Technology
IJST Vol 5 No. 2| July 2026 | ISSN: 2828-7223 (print), ISSN: 2828-7045 (online), Page 248-260

Heart Disease Prediction Using Logistic Regression and K-Nearest Neighbor: A Comparative Study of Classification Algorithm Performance

Yan Risa Aspi Sururi[1*], M. Asadullah Al Ghozi[2]
Universitas Gunadarma, Indonesia
Universitas Indonesia, Indonesia

Article History
Received : June 19, 2026
Revised : July 26, 2026
Accepted : July 28, 2026
Published : July 29, 2026

Corresponding author*:
aspie@staff.gunadarma.ac.id

Cite This Article [*APA Style*]:
Sururi, Y. R. A., & Ghozi, M. A. A. (2026). Heart Disease Prediction Using Logistic Regression and K-Nearest Neighbor: A Comparative Study of Classification Algorithm Performance. *International Journal Science and Technology*, *5*(2), 248–260.

DOI:
https://doi.org/10.56127/ijst.v5i2.2092

Abstract: Heart disease remains one of the leading causes of mortality worldwide, highlighting the importance of accurate and timely prediction models to support early clinical decision-making. **Objective:** This study aims to compare the predictive performance of Logistic Regression and K-Nearest Neighbor (KNN) algorithms for heart disease classification and to identify the most appropriate model for early screening. **Methodology:** A quantitative comparative research design was employed using the Cleveland Heart Disease dataset from the UCI Machine Learning Repository, consisting of 303 patient records. The data were divided into training and testing sets using an 80:20 split. Logistic Regression was developed using backward stepwise selection, while KNN used standardized numerical variables with K = 15. Model performance was evaluated using accuracy, sensitivity, specificity, and precision. **Findings:** Logistic Regression outperformed KNN by achieving an accuracy of 83.6%, sensitivity of 93.5%, specificity of 73.3%, and precision of 78.4%. In comparison, KNN achieved an accuracy of 80.3%, sensitivity of 89.5%, specificity of 65.2%, and precision of 81.0%. These results indicate that Logistic Regression provides more reliable overall performance, particularly in identifying patients with heart disease. **Implications:** The findings suggest that Logistic Regression is more suitable as a decision-support model for early heart disease screening due to its higher sensitivity, accuracy, and specificity. This model may support healthcare professionals in identifying at-risk patients and reducing the likelihood of missed heart disease cases. **Originality:** The originality of this study lies in its transparent comparison of Logistic Regression and KNN using a standardized benchmark dataset while emphasizing clinically relevant evaluation metrics, particularly sensitivity. This approach provides additional empirical evidence for selecting interpretable machine learning models in clinical prediction tasks.

Keywords: heart disease; logistic regression; K-nearest neighbor; machine learning; clinical decision support.

INTRODUCTION

Cardiovascular disease remains the leading cause of death worldwide, accounting for an estimated 17.9 million deaths annually, or approximately 32% of all global deaths. Among cardiovascular diseases, ischemic heart disease has consistently ranked as the leading cause of mortality both before and after the COVID-19 pandemic (Ministry of Health of the Republic of, 2024a). In Indonesia, the 2023 Indonesian Health Survey (Survei Kesehatan Indonesia/SKI) reported a physician-diagnosed heart disease prevalence of

0.85% of the total population, increasing from 0.5% in 2013 to 1.5% in the 2018 Basic Health Research (Riskesdas). Furthermore, expenditures for cardiovascular diseases under the National Health Insurance (Jaminan Kesehatan Nasional/JKN) reached IDR 22.8 trillion in 2023 (Ministry of Health of the Republic of, 2024b). A particularly concerning finding is the shift in the age distribution of patients toward the productive-age population, with individuals aged 25–34 years representing the largest group of heart disease patients. This trend suggests that lifestyle changes have accelerated the onset of heart disease among younger populations (GoodStats, 2024).

Early detection of heart disease is essential for reducing mortality, as approximately 80% of heart disease cases are preventable through timely detection and intervention (Ministry of Health of the Republic of, 2024a). However, conventional diagnostic procedures, which rely on the manual interpretation of multiple clinical indicators by healthcare professionals, are relatively time-consuming and susceptible to inter-observer variability. Recent advances in machine learning have created opportunities to develop predictive models capable of classifying heart disease risk rapidly, consistently, and in a data-driven manner, thereby serving as clinical decision support tools (Chandra & Verma, 2025). Various classification algorithms have been investigated for this purpose, including Logistic Regression, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest, and Decision Tree. Their predictive performance varies depending on dataset characteristics and the preprocessing techniques employed (Hammoud et al., 2024).

Logistic Regression is widely applied in clinical research due to its simplicity, coefficient interpretability, and strong statistical foundation, making its results relatively easy for healthcare professionals to interpret (Suresh et al., 2024). In contrast, K-Nearest Neighbor (KNN) is a popular benchmark algorithm because of its intuitive and non-parametric approach, although its performance depends heavily on the selection of the optimal K value and the scaling of predictor variables (Suresh et al., 2024). Comparative studies evaluating Logistic Regression and KNN on different heart disease datasets have reported that Logistic Regression outperforms KNN across several evaluation metrics in certain cases. Nevertheless, these findings cannot be generalized, as model performance is substantially influenced by differences in datasets, data partitioning strategies, and preprocessing procedures adopted across studies (Hammoud et al., 2024). This research gap highlights the need to replicate comparative analyses using a widely recognized benchmark clinical dataset, namely the Cleveland Heart Disease dataset from the UCI

Machine Learning Repository, while employing transparent and reproducible preprocessing and evaluation methodologies.

Based on the foregoing discussion, this study aims to: (1) develop a heart disease prediction model using Logistic Regression with stepwise variable selection; (2) develop a heart disease prediction model using K-Nearest Neighbor (KNN) with numerical feature standardization; and (3) compare the performance of both models using accuracy, sensitivity, specificity, and precision to identify the most appropriate model for early screening purposes. Theoretically, this study contributes to the growing body of empirical evidence comparing classification algorithms on a standardized clinical dataset. Practically, its findings may serve as a preliminary reference for developing cost-effective clinical decision support systems in healthcare facilities with limited resources. This study is limited to the Cleveland Heart Disease dataset, consisting of 303 observations, and does not include external validation using an Indonesian patient population. Therefore, the generalizability of the findings to local clinical settings should be interpreted with caution and requires further validation.

RESEARCH METHOD

Mathematically, Logistic Regression models the probability of the positive class (heart disease) using the logit function, as expressed in Equation (1).

Equation 1

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}}$$

Where $P(y = 1 \mid x)$denotes the probability that a patient has heart disease, $b_0$is the intercept, $b_1$to $b_k$are the regression coefficients associated with predictors $x_1$to x_k, and $e$is the base of the natural logarithm. In contrast, the K-Nearest Neighbor (KNN) algorithm classifies each new observation based on the majority class among its *K* nearest neighbors, with the distance between observations calculated using the Euclidean distance, as shown in Equation (2).

Equation 2

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Where $d(x, y)$represents the Euclidean distance between observations $x$and y, $x_i$and $y_i$denote the values of the i-th variable for the respective observations, and $n$is the number of numerical variables used. Both models were evaluated using four performance metrics derived from the confusion matrix, as presented in Equations (3)–(6), where TP (true positive), TN (true negative), FP (false positive), and FN (false negative) represent the classification outcomes on the testing dataset.

Equation 3

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

Equation 4

$$Sensitivity = \frac{TP}{TP + FN}$$

Equation 5

$$Specificity = \frac{TN}{TN + FP}$$

Equation 6

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

This study employed a quantitative comparative research design to evaluate the performance of two classification algorithms, namely Logistic Regression and K-Nearest Neighbor (KNN), in predicting patients' heart disease status. The dataset used was the Cleveland Heart Disease dataset obtained from the UCI Machine Learning Repository (Janosi et al., 1988), consisting of 303 patient observations and 14 clinical variables, as summarized in Table 1. The target variable (*target*) was binary, where a value of 1 indicated the presence of heart disease and 0 indicated its absence. The original class distribution comprised 165 patients (54.5%) diagnosed with heart disease and 138 patients (45.5%) without heart disease, as illustrated in Figure 1.

Tabel 1. Research Variables on Dataset Cleveland Heart Disease Dataset

Variables	Description	Data Types
age	Patient age (year)	Numerical
sex	Sex (1 = male, 0 = female)	Category
cp	Chest pain type (4 categories)	Category
trestbps	Resting blood pressure (mmHg)	Numerical
chol	Serum cholesterol level (mg/dL)	Numerical
fbs	Fasting blood sugar > 120 mg/dL	Category
Restecg	Resting electrocardiographic results	Numerical
thalach	Maximum heart rate achieved	Numerical
exang	Exercise-induced angina (1 = yes, 0 = no)	Category

Variables	Description	Data Types
oldpeak	ST-segment depression induced by exercise	Numerical
slope	Slope of the peak exercise ST segment	Numerical
ca	Number of major vessels obstructed (0–3)	Numerical
thal	Thalassemia status (normal/fixed defect/reversible defect)	Numerical
target	Heart disease status (1 = yes, 0 = no)	Category

Sumber: Janosi et al. (1988), UCI Machine Learning Repository.

Prior to model development, categorical variables stored in numerical format (*sex*, *cp*, *fbs*, *exang*, and *target*) were converted into factor variables. The dataset was subsequently partitioned into training data (80%, 242 observations) and testing data (20%, 61 observations) using random sampling with a fixed random seed to ensure reproducibility. The class distributions in the training set (55.4% positive cases) and testing set (50.8% positive cases) indicated a reasonably balanced data split.

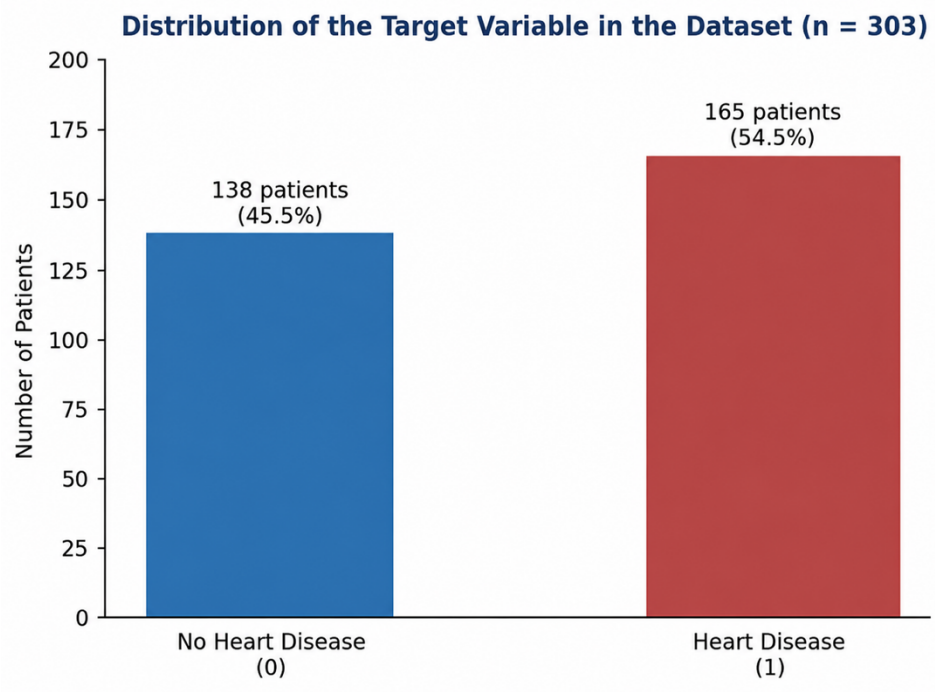

Figure 1. Distribution of Target Variables in the Dataset (n = 303)
Source: Data processed by authors.

The Logistic Regression model was developed using the **glm** function with the binomial family, initially including all predictor variables. The model was then simplified through backward stepwise selection based on the Akaike Information Criterion (AIC) to obtain the most parsimonious model containing only statistically significant predictors. The KNN model was constructed using only nine purely numerical variables (*age*, *trestbps*, *chol*, *restecg*, *thalach*, *oldpeak*, *slope*, *ca*, and *thal*), as the algorithm relies on Euclidean

distance, which is not appropriate for categorical variables encoded as numeric values. All numerical variables were standardized using *z*-score normalization before model training to prevent variables with larger measurement scales from dominating the distance calculations. The means and standard deviations computed from the training data were subsequently applied to standardize the testing data in order to prevent data leakage. A value of $K = 15$ was selected based on the square root of the number of training observations ($\sqrt{242} \approx 15.6$), a commonly used heuristic in KNN applications.

The predictive performance of both models was evaluated using the confusion matrix based on four primary metrics: accuracy (the proportion of correctly classified observations), sensitivity or recall (the ability to correctly identify patients with heart disease), specificity (the ability to correctly identify patients without heart disease), and precision (the proportion of positive predictions that are correct). Sensitivity was considered the primary evaluation metric in the context of heart disease screening because failing to identify patients who truly have the disease (false negatives) may delay critical medical intervention. All data preprocessing and model development were performed using the R programming language. The dplyr and ggplot2 packages were used for data manipulation and visualization, caret and class for model development and evaluation, and MLmetrics and gtools for supplementary performance metric calculations.

The methodological limitations of this study include the relatively small sample size (303 observations), which may affect the stability of model estimates; the use of a single train-test split without repeated cross-validation (e.g., *k*-fold cross-validation), which could provide more robust estimates of predictive performance; and the absence of external validation using populations beyond the Cleveland dataset. The dataset employed in this study is publicly available and had been fully anonymized by the data provider prior to its release. Therefore, no additional ethical approval was required for its use in this research.

RESULT

Logistic Regression Model Results and Evaluations

The backward stepwise selection procedure produced a final model comprising nine significant predictors: sex, chest pain type, resting blood pressure, maximum heart rate achieved, exercise-induced angina, oldpeak**,** slope**,** the number of major vessels colored by fluoroscopy (ca), and thalassemia status (thal), with an Akaike Information Criterion (AIC)

value of 185.25. Figure 2 presents the statistically significant predictor coefficients ($p < 0.05$).

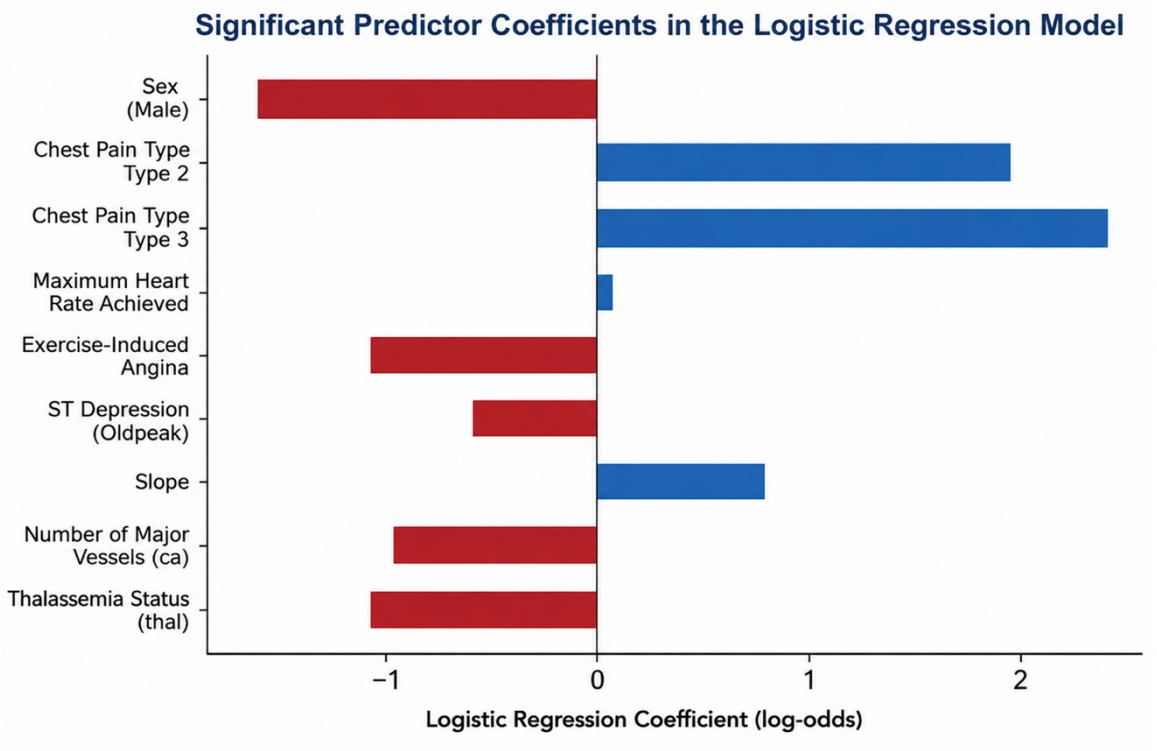

Figure 2. Significant Predictor Coefficients in the Logistic Regression Model
Source: Data processed by authors.

The largest positive coefficients were observed for chest pain type category 3 ($\beta = 2.421; p = 0.001$) and category 2 ($\beta = 1.985; p < 0.001$), indicating that patients with these chest pain types were substantially more likely to be classified as having heart disease than those in the reference category. Conversely, **ca** (number of major vessels colored by fluoroscopy) and thal (thalassemia status) exhibited significant negative coefficients ($\beta = -0.927$ and $\beta = -1.021$, respectively). This finding is clinically consistent with the variable coding adopted in the dataset, in which higher values of these variables are associated with lower log-odds under the target coding scheme used in the present model.

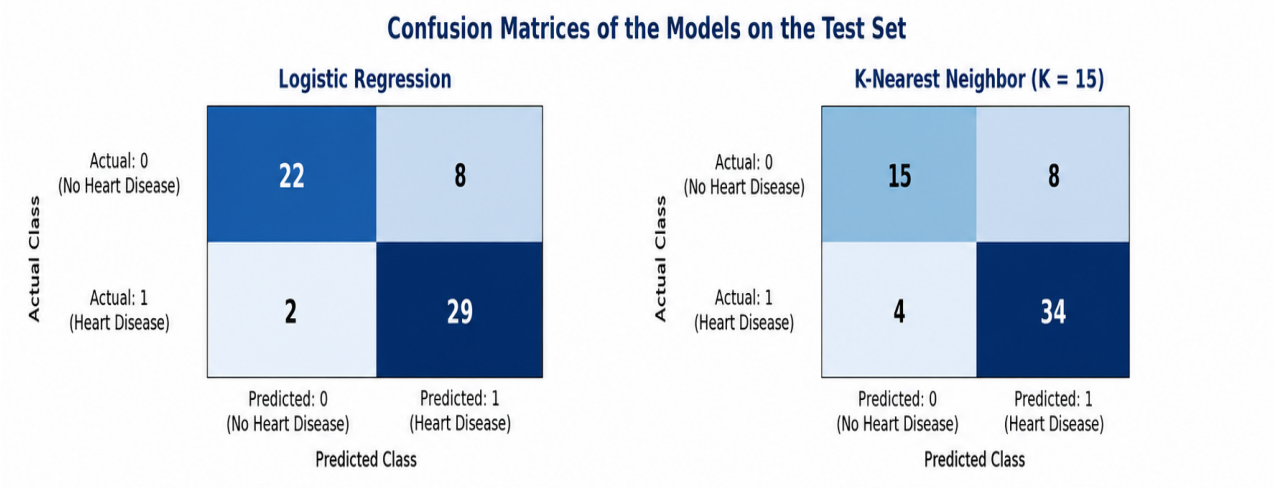

Figure 3. Confusion Matrix of Logistic Regression and KNN Models
Source: Data processed by authors.

Evaluation on the testing dataset produced the confusion matrices shown in Figure 3. The Logistic Regression model correctly classified 22 of the 24 patients without heart disease and 29 of the 37 patients with heart disease. In comparison, the K-Nearest Neighbor (KNN) model correctly classified 15 of the 19 patients without heart disease and 34 of the 42 patients with heart disease in its respective testing dataset.

Based on these confusion matrices, Table 2 summarizes the four primary evaluation metrics for both models, while Figure 4 provides a side-by-side comparison of their predictive performance.

DISCUSSION

The results indicate that the Logistic Regression model outperformed the K-Nearest Neighbor (KNN) model on three of the four evaluation metrics, achieving higher accuracy (83.6% vs. 80.3%), sensitivity (93.5% vs. 89.5%), and specificity (73.3% vs. 65.2%). Conversely, the KNN model demonstrated a slightly higher precision (81.0% vs. 78.4%). The superior sensitivity of Logistic Regression represents the most clinically significant finding, as this metric reflects the model's ability to correctly identify patients who truly have heart disease. High sensitivity minimizes the risk of false negatives, a condition in which patients with heart disease are incorrectly classified as healthy and may consequently fail to receive timely medical treatment. These findings are consistent with Pareek et al. (2025), who reported that Logistic Regression achieved higher predictive accuracy than competing methods in similar classification tasks, and with Suresh et al. (2024), who

highlighted the widespread use of Logistic Regression in clinical research due to its simplicity and interpretability.

Table 2. Comparison of Logistic Regression and KNN Model Performance

Metrics	Logistic Regressions	KNN (K=15)
Accuracy	0,836	0,803
Sensitivity (Recall)	0,935	0,895
Spesificity	0,733	0,652
Precision	0,784	0,810

Source: Data processed by authors.

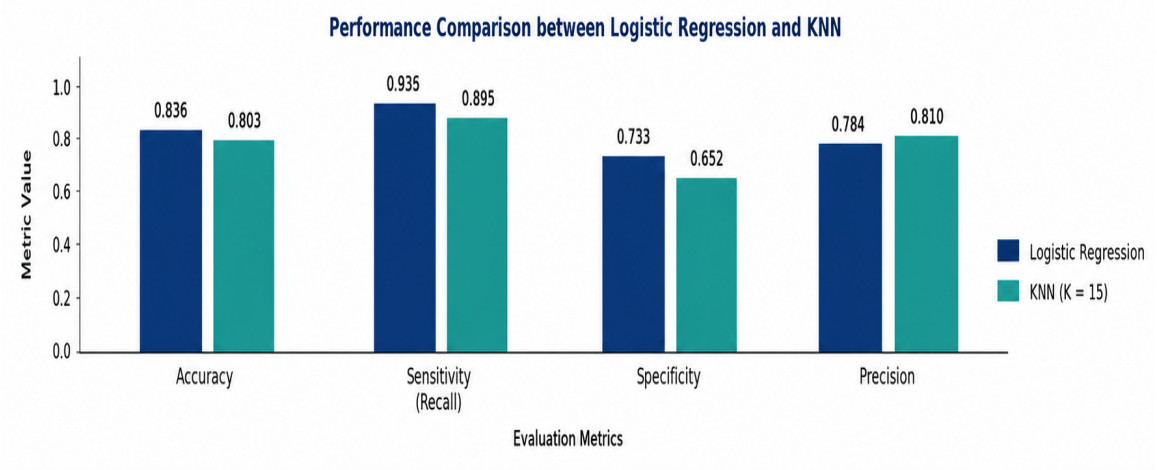

Figure 4 Comparison of Logistic Regression and KNN Performance
Source: Data processed by authors.

The slightly lower performance of the KNN model across most evaluation metrics can be attributed to the characteristics of the algorithm, which is highly dependent on the selection of the optimal K value and the local density of observations within the feature space. With only nine predictor variables and a training set consisting of 242 observations, the resulting decision boundaries were likely less optimal than those generated by the parametric Logistic Regression model, which simultaneously incorporated all nine significant predictors, including categorical variables such as chest pain type, one of the strongest predictors identified in this study.

Table 3. Comparison of Research Results with Related Studies

Study	Dataset	Main Findings
This study (2026)	Cleveland (n=303)	Logistic regression outperforms KNN (83.6% accuracy; 93.5% sensitivity)
(Pareek et al., 2025)	Combination of clinical dataset	Logistic regression and KNN improve prediction accuracy compared to Naive Bayes
(Hammoud et al., 2024)	Cleveland, Statlog, Hungary (n=1189)	Seven algorithms compared; performance varies depending on feature selection technique
(Islam et al., 2025)	Open heart disease dataset	KNN achieves 93% macro accuracy with cross-validation and more features
(Chang et al., 2022)	Cleveland Heart Disease	K-Nearest Neighbor achieved the highest initial score of 87%, followed by Random Forest at 84%
(Bilal et al., 2025)	MIT-BIH Arrhythmia Dataset (109,446 ECG signal instances)	Utilizing SMOTE for dataset balancing, the hybrid model achieved a superior testing accuracy of 99.70%, significantly outperforming standalone models
(Nagavelli et al., 2022)	Cleveland and Statlog datasets from the UCI	XGBoost as the top-performing algorithm with an accuracy of 95.9%
(Teja & Rayalu, 2025)	Integrated UCI Dataset	XGBoost and Bagged Trees reached the highest overall accuracy of 93%. While Random Forest exhibited the most stability during 10-fold cross-validation (94% accuracy)

Source: Processed from various related studies.

Table 3 summarizes the findings of the present study alongside several comparable studies that evaluated Logistic Regression and K-Nearest Neighbor (KNN) algorithms using different heart disease datasets. This comparison provides broader context regarding both the consistency and the variability of findings reported in the existing literature.

Previous comparative studies using different heart disease datasets have produced mixed findings. Some reported that KNN achieved macro accuracy of up to 93% when combined with cross-validation and a larger number of features (Islam et al., 2025), whereas others confirmed the superiority of Logistic Regression on datasets with class distributions and feature dimensions similar to those used in the present study (Pareek et al., 2025). These discrepancies underscore that the relative performance of the two algorithms is highly dependent on dataset characteristics, highlighting the continued importance of replicating comparative studies using benchmark datasets such as the Cleveland Heart Disease dataset to strengthen the empirical evidence base.

Logistic Regression is favored in medical domains due to its interpretability and strong statistical foundation (Teja & Rayalu, 2025). However, some literature indicates that KNN can reach high macro-accuracy (93%) when combined with cross-validation and larger

feature sets, though it remains highly dependent on the choice of the K value (Teja & Rayalu, 2025). The adoption of ensemble methods has further enhanced diagnostic precision beyond baseline classifiers. Algorithms such as XGBoost, Random Forest (RF), and Bagged Trees consistently outperform individual models ((Nagavelli et al., 2022; Teja & Rayalu, 2025) For instance, XGBoost has been reported to reach accuracies between 93% and 95.9% (Nagavelli et al., 2022; Teja & Rayalu, 2025). Additionally, Random Forest has demonstrated remarkable stability during 10-fold cross-validation, maintaining a 94% accuracy rate, while ensemble models overall provide superior ROC-AUC values (up to 95%) (Teja & Rayalu, 2025).

Recent breakthroughs in hybrid deep learning represent the current state-of-the-art for heart disease diagnosis, particularly when processing ECG signals (Bilal et al., 2025). A proposed hybrid model combining Bidirectional Long-Short-Term Memory (BLSTM) and Bidirectional Gated Recurrent Units (BGRU) known as BLSTM-BGRU demonstrates extreme precision by capturing both long-range dependencies and short-term variations in heart rate data (Bilal et al., 2025). This architecture reduces trainable parameters and overcomes vanishing gradient issues, achieving an unprecedented testing accuracy of 99.70% on balanced datasets (Bilal et al., 2025)

Crucial to the success of these models is the preprocessing stage, which addresses data quality issues like class imbalance and outliers. The use of SMOTE (Synthetic Minority Over-sampling Technique) is widely documented for balancing clinical datasets to prevent model overfitting (Bilal et al., 2025; Nagavelli et al., 2022). Furthermore, advanced outlier elimination techniques like DBSCAN and the application of SMOTE-ENN for data cleaning have been shown to further refine training data (Nagavelli et al., 2022)

CONCLUSION

This study demonstrates that Logistic Regression provides superior predictive performance compared with the K-Nearest Neighbor (KNN) algorithm for heart disease classification using the Cleveland Heart Disease dataset. Logistic Regression achieved higher accuracy (83.6%), sensitivity (93.5%), and specificity (73.3%), whereas KNN showed only a marginal advantage in precision. From a clinical perspective, the higher sensitivity of Logistic Regression represents the most important finding because it reduces the likelihood of false-negative predictions, thereby supporting more reliable early heart disease screening. Furthermore, chest pain type, the number of major vessels identified by

fluoroscopy (ca), and thalassemia status (thal) were identified as the most influential predictors in the final Logistic Regression model, highlighting their relevance in heart disease prediction.

The primary scientific contribution of this research lies in providing an updated comparative evaluation of Logistic Regression and KNN using a standardized benchmark dataset with transparent preprocessing and reproducible modelling procedures. Unlike many previous studies that focused mainly on predictive accuracy, this study emphasizes clinically meaningful evaluation metrics, particularly sensitivity, specificity, and precision, to support decision-making in early disease screening. The findings strengthen empirical evidence that interpretable statistical models such as Logistic Regression remain competitive for clinical prediction tasks, particularly in healthcare settings where model transparency and ease of interpretation are essential for practical implementation.

This study has several limitations. First, the analysis was conducted using only the Cleveland Heart Disease dataset, consisting of 303 observations, which may limit the generalizability of the findings to broader patient populations. Second, model evaluation relied on a single train-test split without repeated cross-validation, potentially affecting the robustness of the reported performance estimates. Third, the study did not include external validation using Indonesian clinical data or compare Logistic Regression and KNN with more advanced ensemble and deep learning methods such as Random Forest, XGBoost, or hybrid neural network architectures. Future research is therefore recommended to employ larger and more diverse datasets, implement k-fold cross-validation and external validation, incorporate feature engineering and class-balancing techniques, and compare conventional machine learning algorithms with state-of-the-art ensemble and deep learning approaches to develop more robust and clinically applicable heart disease prediction models.

REFERENCES

Bilal, H., Muhammad, Y., Ullah, I., Garg, S., Choi, B. J., & Hassan, M. M. (2025). Identification and Diagnosis of Chronic Heart Disease: A Deep Learning-Based Hybrid Approach. *Alexandria Engineering Journal*, *124*, 470-483. https://doi.org/10.1016/j.aej.2025.03.025

Chandra, S., & Verma, P. (2025). A Comprehensive Review of Machine Learning for Heart Disease Prediction: Challenges, Trends, Ethical Considerations, and Future Directions. *Frontiers in Digital Health*, *7*, 1-18.

Chang, V., Bhavani, V. R., Xu, A. Q., & Hossain, M. A. (2022). An Artificial Intelligence Model for Heart Disease Detection Using Machine Learning Algorithms. *Healthcare Analytics*, *2*, 100016. https://doi.org/10.1016/j.health.2022.100016

GoodStats. (2024). Pasien Jantung di Indonesia Didominasi Usia Produktif. https://data.goodstats.id/statistic/pasien-jantung-di-indonesia-didominasi-usia-produktif-79yo9

Hammoud, A., Karaki, A., Tafreshi, R., Abdulla, S., & Wahid, M. (2024). Coronary Heart Disease Prediction: A Comparative Study of Machine Learning Algorithms. *Journal of Advances in Information Technology*, *15*(1), 27-32. https://doi.org/10.12720/jait.15.1.27-32

Islam, M. R., Rahman, S., & Chowdhury, F. (2025). Comparative Analysis of Heart Disease Prediction Using Logistic Regression, SVM, KNN, and Random Forest with Cross-Validation for Improved Accuracy. *Scientific Reports*, *15*, 13444. https://doi.org/10.1038/s41598-025-13444-x

Ministry of Health of the Republic of, I. (2024a). *Indonesian Health Profile 2024*.

Ministry of Health of the Republic of, I. (2024b). *Indonesian Health Survey (SKI) 2023*.

Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine Learning Technology-Based Heart Disease Detection Models. *Journal of Healthcare Engineering*, *2022*, 7351061. https://doi.org/10.1155/2022/7351061

Pareek, P., Kumari, N., & Yadav, P. (2025). Heart Disease Prediction Using Machine Learning Algorithms: A Comparative Study of Logistic Regression and KNN. *International Journal of Research - GRANTHAALAYAH*, *13*(1), 141-154. https://doi.org/10.29121/granthaalayah.v13.i1.2025.6123

Suresh, N., Thomas, B., & Joseph, J. (2024). Bibliometric Analysis and Visualization of Scientific Literature on Heart Disease Classification Using a Logistic Regression Model. *Cureus*, *16*(6), e63337. https://doi.org/10.7759/cureus.63337

Teja, M. D., & Rayalu, G. M. (2025). Optimizing Heart Disease Diagnosis with Advanced Machine Learning Models: A Comparison of Predictive Performance. *BMC Cardiovascular Disorders*, *25*, 212. https://doi.org/10.1186/s12872-025-04627-6