International Journal of Science and Mathematics Education
Volume 1, Nomor 4, Tahun 2024
e-ISSN: 3062-7214; p-ISSN: 3062-7206, Hal 19-38
DOI ; https://doi.org/10.62951/ijsme.v1i4.96
Available online at: https://international.arimsi.or.id/index.php/IJSME

Optimizing Heart Disease Prediction : A Comparative Study of Machine Learning Models Using Clinical Data

Budiman[1]*, Nur Alamsyah[2], Elia Setiana[3], Valencia Claudia Jennifer Kaunang[4], Syahira Putri Himmaniah [5]
[1,2,3,4,5] Universitas Informatika dan Bisnis Indonesia, Indonesia

Address: Jl. Soekarno-Hatta No.643, Sukapura, Kiaracondong District, Bandung City, West Java 40285
*Corresponding Author: budiman@unibi.ac.id**

***Abstrac** : Cardiovascular disease is a leading cause of death globally, necessitating effective predictive systems. This research aims to analyze the effectiveness of various machine learning (ML) models—Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN)—in predicting heart disease using publicly available health data. The study involved pre-processing data, training models, and evaluating them using accuracy, precision, recall, F1-score, and G-Mean metrics. The results show that KNN is the most reliable model, with the highest accuracy of 92%. Significant health features were identified, such as chest pain type and maximum heart rate. The study contributes to improving clinical decision support systems by identifying optimal ML models for heart disease prediction..*

***Keywords:** Heart disease prediction, machine learning, Logistic Regression, K-Nearest Neighbors, health data*

1. INTRODUCTION

Cardiovascular disease is the leading cause of death worldwide according to the World Health Organisation (WHO), with 17.9 million deaths annually. In addition, the risk of heart disease increases due to harmful behaviours that can lead to overweight, obesity, hypertension, hyperglycaemia and high cholesterol (*Cardiovascular Diseases*, n.d.). Furthermore, healthcare data also has great potential to be used in the development of effective health prediction systems to improve heart disease prevention efforts (Ismail et al., 2020). Diagnosis of heart disease is a challenge that can be overcome through computerised estimation of disease severity, so that follow-up actions can be taken more easily (Gokulnath & Shantharajah, 2019). In addition, early diagnosis of heart disease is crucial as it can improve the appropriateness of treatment and also provide faster diagnostic recommendations from clinical experts (Sekar et al., 2022). Thus, timely diagnosis plays a crucial role in reducing further health risks while preventing heart attacks (Dutta et al., 2020). Lastly, electrocardiogram (ECG) is widely used in the clinical diagnosis of heart disease as it is able to fully represent the electrical activity of the heart on the surface of the human body(Alarsan & Younes, 2019).

Pattern prediction to prevent and control diseases is a significant challenge and an important need in the medical domain. Healthcare data can be used to develop health prediction systems that are effective in improving heart disease prevention. Big data about healthcare, such as patient records, clinical records, diagnoses, past illnesses of parents and families,

Received Oktober 09, 2024; Revised November 21, 2024; Accepted Desember 10, 2024; Published Desember 12, 2024

hospitals, as well as scan results, can help in the stage of disease identification and prediction (Ismail et al., 2020). Heart disease prediction is very useful because it can help health practitioners make more accurate decisions about the patient's health condition. Therefore, the use of machine learning (ML) is a relevant solution to reduce and understand the symptoms associated with heart disease (Gárate-Escamila et al., 2020). Moreover, data mining plays a very important role in the information extraction process, by identifying hidden sequences, performing regression and classification, building analytical models, performing clustering, and representing the mined results using various visualisation and presentation techniques (Rahman et al., 2021; Repaka et al., 2019; Sireesha, 2020). To address this issue, many approaches based on data mining techniques have been proposed in recent years to assist healthcare professionals in diagnosing heart disease. Heart disease prediction systems that utilise data mining techniques can assist clinicians in providing more accurate predictions based on patients' clinical data (Gárate-Escamila et al., 2020; Le et al., 2018).

This research aims to analyse the effectiveness of various machine learning models, such as Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN), in predicting heart disease using health data. This research also seeks to develop a heart disease prediction system by utilising publicly available datasets, by analysing various features such as chest pain type, cholesterol level, and electrocardiogram results. In addition, this research compares the performance of machine learning algorithms based on accuracy, precision, recall, F1-score, and G-Mean metrics in the context of heart disease prediction. Another objective of this study was to identify the most significant health features in influencing heart disease prediction and determine their role in the development of clinical decision support systems. Finally, this study evaluates the potential application of machine learning-based heart disease early detection systems in clinical environments to improve diagnosis and more targeted treatment.

2. LITERATURE REVIEW

Logistic regression is a classification model used to predict the probability of an event, usually in binary cases (e.g., heart disease or not). Although called ‘regression’, this model is used for classification. The model uses a sigmoid function to map linear inputs into probabilities. Logistic Regression works by finding coefficients that maximise the likelihood of prediction on the training data. The advantages of Logistic Regression are its simplicity and good interpretability, as each coefficient can be interpreted as the logarithmic effect of changing the input variable on the output probability (Anjum et al., 2024).

Random Forest is a decision tree-based ensemble model that combines multiple decision trees to make more accurate and stable predictions. Each tree in Random Forest is trained using a random subset of data and features, which reduces overfitting and improves generalisation (*Recursion Enhanced Random Forest With an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform | IEEE Journals & Magazine | IEEE Xplore*, n.d.).

Naive Bayes is a probabilistic model based on Bayes' Theorem with the assumption that the features are independent. This model is often used for text classification and problems involving categorical data. Naive Bayes is very fast as it only needs to calculate the probability of each feature separately, making it suitable for large datasets. However, the assumption of independence between features is often unrealistic, so performance may degrade on more complex datasets (Repaka et al., 2019).

Support Vector Classifier (SVC) works by finding a hyperplane that separates two classes in the feature space. The goal is to find the hyperplane with the largest margin (distance between data from two classes) so that the model can better predict new data. SVC maximises the margin between two classes by solving an optimisation problem involving Lagrange multipliers. For data that cannot be linearly separated, a kernel trick is used to map the data to a higher dimension, where a hyperplane can be formed. The advantage of SVC is its ability to handle high-dimensional data and remain effective on small datasets. However, if the amount of data is very large, SVC can be slow (Ali et al., 2019; Gokulnath & Shantharajah, 2019, 2019).

K-Nearest Neighbors (KNN) is a classification method that works based on distance. When new data is provided, the model searches for the K nearest neighbours (training data) and classifies them based on the majority class of the neighbours. KNN is very intuitive and does not require an explicit training process. The downside is that KNN can be slow when the amount of data is very large, because every time a prediction is made, the model has to calculate the distance to all the training data (Rahman et al., 2021).

3. METHODOLOGY

The methodology illustrated in Figure 1 outlines the steps involved in developing the proposed system for heart disease prediction. It begins with a literature study to understand existing approaches and identify research gaps.

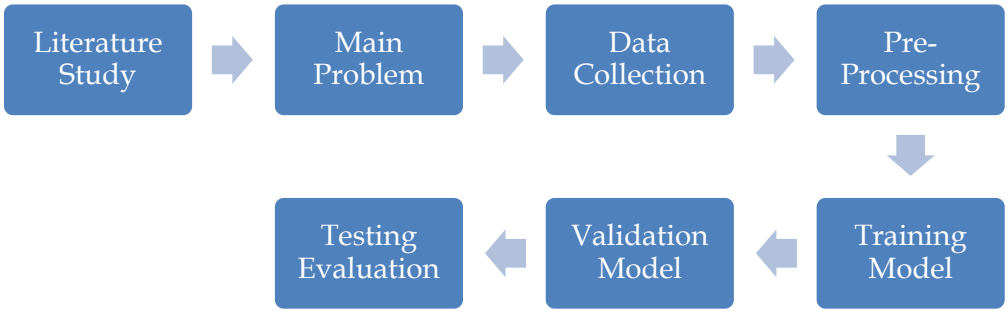

Figure 1. Proposed Method

Literature Study

In this study, we draw on relevant literature on heart disease prediction using machine learning. Previous research shows that models such as Logistic Regression, Random Forest, Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN) have been widely used in health classification. In addition, pre-processing such as data normalisation, missing value handling, and feature selection have been shown to contribute significantly in improving model accuracy.

Main Problem

The main problem to be solved is to build a model that can predict whether a patient has heart disease based on health features such as age, gender, blood pressure, cholesterol, and others. The main objective of this study is to minimise the prediction error and improve the performance of the model in detecting patients with heart disease. The main objective of this study is to minimise the prediction error and improve the performance of the model in detecting patients with heart disease.

Data Collection

We used the publicly available Heart Disease dataset from kaggle, which consists of several features related to the patient's health condition and the target variable (heart disease or not). The Heart Disease dataset has 303 records and 14 features related to risk factors and characteristics that can affect a person's likelihood of developing heart disease. The following is an explanation of the features in this dataset, displayed in the form of table 1.

Table 1. Heart Disease Feature

Fitur	Description
age	Patient's age in years
sex	Patient's gender (1 = male, 0 = female)
cp	Type of chest pain (0: none, 1: atypical angina, 2: typical angina, 3: non-angina)
trestbps	Resting blood pressure in mm Hg (millimeters of mercury)

chol	Serum cholesterol level in mg/dl (milligrams per deciliter)
fbs	Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
restecg	Electrocardiographic results at rest (0 = normal, 1 = has abnormalities, 2 = left ventricular hypertrophy)
thalach	Maximum heart rate achieved
exang	Exercise-induced angina (1 = yes, 0 = no)
oldpeak	Exercise-induced ST depression relative to resting state
slope	The slope of the ST segment during exercise (0 = decreased, 1 = flat, 2 = increased)
ca	Number of large blood vessels stained by fluoroscopy (values from 0 to 3)
thal	Thalassemia status (1 = normal, 2 = permanent disability, 3 = curable disability)
target	Diagnosis of heart disease (1 = yes, 0 = no)

Pre-Processing

The pre-processing stage in this research methodology is an important step to prepare the data for use in machine learning models. This process starts with handling missing values, where incomplete data is cleaned or accounted for so as not to affect the model training results. Next, data normalization is performed to scale the various features, such as age, blood pressure, and cholesterol levels, so that each feature has equal weight in the training process. This normalization technique is important to ensure that the model can optimally understand the data patterns without being affected by the scaling differences between features. After that, irrelevant or redundant features are removed through feature selection, with the aim of improving model efficiency and reducing computational complexity. The final step in pre-processing is to separate the dataset into training set and testing set to ensure that the model is trained and tested independently. By doing good pre-processing, it is expected that machine learning models can learn the data more effectively and produce more accurate predictions.

Training Model

In the model training stage of this research methodology, we used a pre-processed heart disease dataset. This dataset is divided into two parts, 80% for training set and 20% for testing set. The machine learning models used in this study include Logistic Regression, Random Forest, Naive Bayes, Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN). Each model is trained using a training set to learn the patterns and characteristics found in the data, especially in health features such as blood pressure, cholesterol, and maximum heart rate. In this process, optimization techniques such as hyperparameter adjustment are applied to improve the accuracy of the model. After the training process is complete, the model is evaluated using cross-validation to minimize overfitting and ensure that the model can make

reliable predictions when applied to new data. The results of this training are then used as a basis to compare the performance of each model in predicting heart disease.

Validation Model

At this stage, model validation is carried out, one of which uses Cross-validation. Cross-validation is an important method used in research to evaluate the performance of prediction models by dividing data into several subsets or folds. This technique aims to prevent overfitting, which is when the model fits the training data so well that it is less able to generalize to new data. In the context of machine learning research, cross-validation is applied at the model evaluation stage after data is collected, processed, and a prediction model is built. One commonly used form of cross-validation is k-fold cross-validation, where the data is divided into k parts, and the model is trained on k-1 parts, while the remaining parts are used for testing. This process is repeated k times so that each piece of data is used once as a test set. The results of all iterations are then averaged to get an idea of the overall performance of the model.

Testing Evaluation

Once the model validation process is complete, the next important step is to evaluate the performance of the model using the test set, which is data that has not been used during training. This evaluation gives an idea of the extent to which the model is able to make accurate predictions on new data. Some commonly used evaluation metrics include accuracy, precision, recall, and F1-score. Accuracy measures the proportion of correct predictions out of the overall data, providing an overview of the model's overall performance. Precision focuses on the model's ability to identify true positive predictions (e.g., correctly predicting patients who actually have heart disease), reducing false positive errors. Recall, on the other hand, measures how well the model can detect all positive cases, by minimizing errors in identifying patients who actually have the disease but were not detected (false negatives). F1-score combines precision and recall into a single value, which is particularly useful in unbalanced data sets, to provide a balance between the model's ability to accurately predict positive cases and detect all positive cases. Evaluation with this metric allows for a more in-depth analysis of the model's performance in different aspects of classification, looking not only at overall accuracy, but also how the model handles false predictions.

4. RESEARCH RESULT

Pre-Processing

Table 2 displays the top five and bottom five data from the heart disease dataset, where each row represents individuals with various relevant health attributes. The first attribute is age which varies between 37 to 68 years. Gender is expressed with a value of 1 for male and 0 for female, indicating that there is a mix of genders among these individuals. The type of chest pain (cp) ranged from 0 (no chest pain) to 3 (severe chest pain), indicating that the type of chest pain varied among individuals.

Resting blood pressure (trestbps) ranged from 120 to 145 mmHg, while cholesterol varied from 204 mg/dl to 409 mg/dl. Some individuals had high fasting blood sugar levels (fbs), expressed as a value of 1, while others were normal (value 0). Electrocardiography (restecg) results showed some individuals were normal (value 0) and some were abnormal (value 1). The maximum heart rate (thalach) varied from 141 bpm to 187 bpm.

There were some individuals who experienced angina during exercise (exang) with a value of 1, while others did not (value 0). ST depression values after exercise (oldpeak) showed variation from 0.0 to 2.3, indicating potential cardiac ischemia. ST segment slope also showed variation, with values from 0 to 2. The number of large blood vessels (ca) seen on fluoroscopy ranges from 0 to 1, and the thalassemia condition (thal) shows values from 1 to 2.

Finally, the target column indicates whether the individual has heart disease (value 1) or not (value 0). Most of the individuals in this table are diagnosed with heart disease (value 1). The combination of these features is used to analyze the likelihood of heart disease in each individual in this dataset.

Table 2. Heart Disease Dataset

index	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak
0	63	1	3	145	233	1	0	150	0	2.3
1	37	1	2	130	250	0	1	187	0	3.5
2	41	0	1	130	204	0	0	172	0	1.4
3	56	1	1	120	236	0	1	178	0	0.8
4	57	0	0	120	354	0	1	163	1	0.6
298	57	0	0	140	241	0	1	123	1	0.2
299	45	1	3	110	264	0	1	132	0	1.2
300	68	1	0	144	193	1	1	141	0	3.4
301	57	1	0	130	131	0	1	115	1	1.2
302	57	0	1	130	236	0	0	174	0	0

index	oldpeak	slope	ca	thal	target
0	2.3	0	0	1	1
1	3.5	0	0	2	1
2	1.4	2	0	2	1
3	0.8	2	0	2	1
4	0.6	2	0	2	1
298	0.2	1	0	3	0
299	1.2	1	0	3	0
300	3.4	1	2	3	0
301	1.2	1	1	3	0
302	0	1	1	2	0

Table 3 shows descriptive statistics from the Heart Disease dataset providing important information about the characteristics of patients diagnosed with heart disease. The average age of patients in this dataset was about 54.37 years with considerable variation, covering the age range from 29 to 77 years. In terms of gender, the majority of patients are male, with about 68.32% of the total population. The type of chest pain experienced by patients also varies, with the majority experiencing mild to moderate type chest pain. The average resting blood pressure was 131.62 mm Hg, while the average cholesterol level reached 246.26 mg/dl, indicating that many patients had fairly high cholesterol levels.

Table 3. descriptive statistics

feature	count	mean	std	min	max	25%	50%	75%
age	303	54.37	9.08	29	77	47.5	55	61
sex	303	0.68	0.47	0	1	0	1	1
cp	303	0.97	1.03	0	3	0	1	2
trestbps	303	131.62	17.54	94	200	120	130	140
chol	303	246.26	51.83	126	564	211	240	274.5
fbs	303	0.15	0.36	0	1	0	0	0
restecg	303	0.53	0.53	0	2	0	1	1
thalach	303	149.65	22.91	71	202	133.5	153	166
exang	303	0.33	0.47	0	1	0	0	1
oldpeak	303	1.04	1.16	0	6.2	0	0.8	1.6
slope	303	1.40	0.62	0	2	1	1	2
ca	303	0.73	1.02	0	4	0	0	1
thal	303	2.31	0.61	0	3	2	2	3
target	303	0.54	0.50	0	1	0	1	1

Most patients did not have high fasting blood sugar, with an average value for fasting blood sugar features of 0.15. Resting ECG results showed that more than half of the patients

had normal or slightly abnormal results. The maximum heart rate that patients achieve is 149.65 beats per minute on average, and about 32.67% of patients experience exercise-induced angina. The average ST segment depreciation value was 1.04, which indicates the varying degree of ST segment depression among patients.

In addition, the slope of the ST segment during exercise mostly shows a flat or ascending slope, while the average number of large colored blood vessels is 0.73, indicating a small number of affected vessels. The average patient's thalassemia status was in the category of permanent or reversible disability, with an average score of 2.31. Finally, about 54.46% of patients were diagnosed with heart disease, which is indicated by an average target value of 0.54. Overall, these data provide a comprehensive picture of the health factors related to heart disease risk among patients.

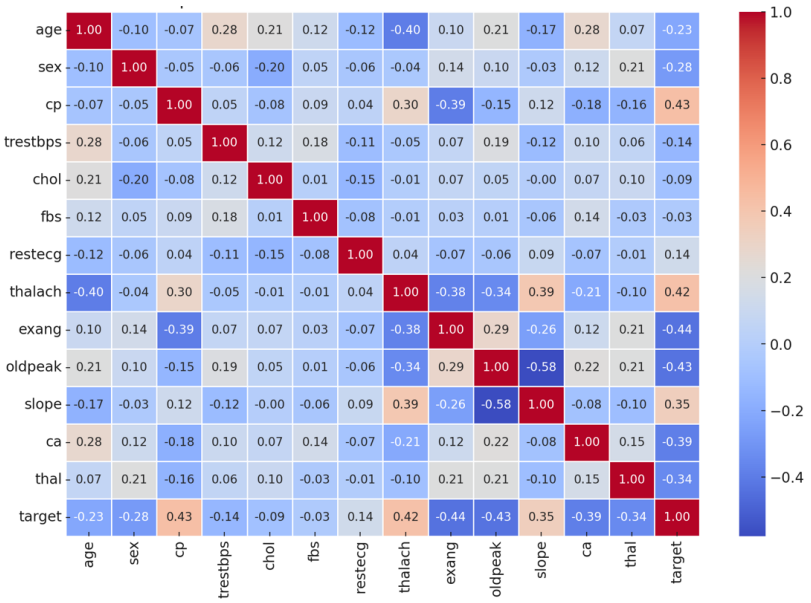

Figure 2. Heatmap

Figure 2 shows a heatmap of the correlation matrix between features in the heart disease dataset, using colors to represent the strength and direction of the correlation. The color scale on the right shows that red indicates a strong positive correlation, blue indicates a strong negative correlation, and neutral colors indicate a weak correlation or no correlation at all. Each cell shows the Pearson correlation coefficient between the two variables, where values close to 1.00 indicate a strong positive relationship, and values close to -1.00 indicate a strong negative relationship. A coefficient close to 0.00 means there is no correlation between the variables.

Several important insights emerged from this heatmap. Some features such as chest pain type (cp) and target variable (with a correlation of 0.43) suggest that certain types of chest pain

are associated with heart disease. Maximum heart rate (thalach) also has a positive correlation of 0.42 with heart disease, while ST segment slope has a correlation of 0.34. These features show a significant association with heart disease. In contrast, other features such as ST depression (oldpeak), exercise-induced angina (exang), and number of major vessels (ca) showed a strong negative correlation with heart disease, with coefficients of -0.43, -0.44, and -0.39, respectively. This suggests that higher values for these features are associated with a lower likelihood of heart disease.

Interestingly, some features such as cholesterol (chol) and fasting blood sugar (fbs) had very little correlation with heart disease, with weak coefficients of -0.09 and -0.05 respectively. Resting electrocardiogram (restecg) results also showed a weak positive association with a correlation of 0.16.

The heatmap also shows strong correlations between some features. For example, the number of major blood vessels (ca) and thalassemia (thal) have a significant correlation of 0.54, and exercise-induced angina (exang) correlates strongly with ST depression (oldpeak) with a coefficient of 0.58. In addition, the relationship between age and maximum heart rate (thalach) showed a negative correlation (-0.40), indicating that older people tend to have a lower maximum heart rate during exercise.

Overall, features such as chest pain type, maximum heart rate, and ST depression showed strong correlations with the likelihood of heart disease and may be important for prediction models. In contrast, features such as cholesterol and fasting blood sugar may be less useful for such models as they show weak or insignificant correlations with the target variables. This heatmap provides a valuable tool for understanding the relationship between features, which is crucial in selecting the most informative predictors in heart disease modeling.

Training Model

The classification models used in this study are Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN) which tested the performance of five main algorithms in predicting heart disease using Python libraries. Data sharing was done with a ratio of 80% for training data and 20% for testing data, ensuring enough data to train the model as well as reliable evaluation on data that the model has never seen.

Validation Model

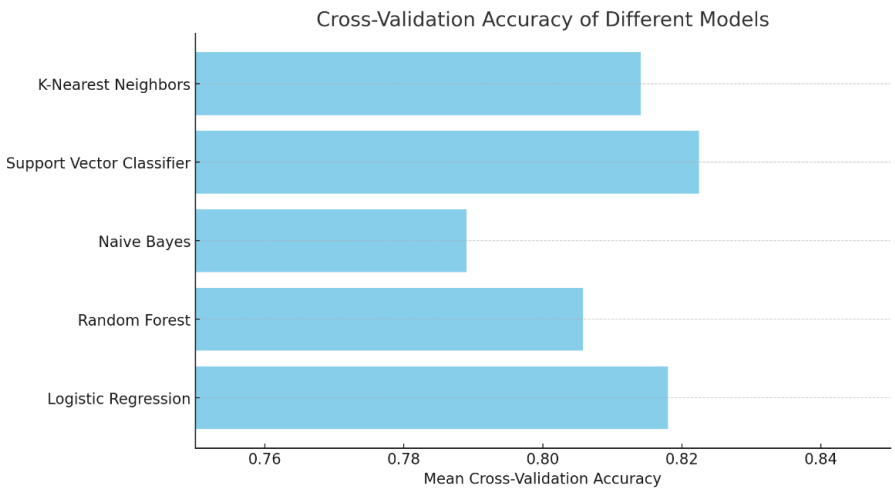

Figure 3. Graph of Cross-validation Results

Figure 3 is a graph of cross-validation results using 5-fold cross-validation showing the performance of several models in predicting heart disease based on the available dataset. Each model is tested using cross-validation techniques to ensure that the prediction results can be generalized well to new unseen data.

Logistic Regression had an average accuracy of 81.8%. This model works under the assumption that there is a linear relationship between input features and output probabilities. This model is often the first choice in binary classification problems due to its simplicity and good interpretability.

Random Forest, with an average accuracy of 80.6%, is slightly lower than Logistic Regression. This model uses an ensemble approach, where a number of decision trees are formed based on random subsets of the data, and the final decision is made through a majority vote. Although flexible and capable of handling more complex data, Random Forest can sometimes suffer from overfitting on data that is too little or not varied enough.

Naive Bayes achieved an average accuracy of 78.9%. Although this is lower than other models, Naive Bayes is still a fast and efficient choice, especially in situations where the assumption of independence between features is valid. Although the independence assumption is often not completely true in many datasets, it is still able to provide quite competitive results.

Support Vector Classifier (SVC) is the second best performing model after KNN, with an average accuracy of 82.24%. SVC works by finding a hyperplane that maximizes the margin between classes in the feature space. This model is particularly effective on datasets that have clear separation margins between classes, as seen in this heart disease data.

K-Nearest Neighbors (KNN) is the best performing model, having an average accuracy of 81.4%. KNN predicts the class of new data based on its proximity to a few nearest neighbors in the feature space. Although this model does not involve an explicit training process, its performance often depends on the selection of the k value (number of neighbors) and the scale of the input features. In this case, KNN managed to provide good prediction results, especially after the data was rescaled with StandardScaler.

Testing Model

Each model was evaluated based on several key metrics, including accuracy, precision, recall, and f1-score for two classes: patients without heart disease (class 0) and patients with heart disease (class 1).

Table 4. Performance of the five classification models

Model	Accuracy	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1-Score (Class 0)	F1-Score (Class 1)
LR	0.85	0.83	0.87	0.86	0.84	0.85	0.86
RF	0.85	0.86	0.85	0.83	0.88	0.84	0.86
NB	0.87	0.84	0.9	0.9	0.84	0.87	0.87
SVC	0.87	0.84	0.9	0.9	0.84	0.87	0.87
KNN	0.92	0.9	0.94	0.93	0.91	0.92	0.92

Logistic regression (LR) has an accuracy of 85%, which shows a fairly good performance in predicting the patient's condition. Precision for class 0 was 0.83, meaning that 83% of predictions indicating that the patient did not have heart disease were correct, while precision for class 1 was 0.87. Recall for both classes was also fairly balanced, with the model detecting 86% of patients without heart disease and 84% of patients with heart disease. The f1 scores were 0.85 and 0.86 respectively, reflecting a good balance between precision and recall.

Logistic Regression (LR) has an accuracy of 85%, which shows a fairly good performance in predicting the patient's condition. Precision for class 0 was 0.83, meaning 83% of the predictions indicating that the patient did not have heart disease were correct, while precision for class 1 reached 0.87. Recall for both classes is also quite balanced, with the model detecting 86% of patients without heart disease and 84% of patients with heart disease. The f1-score values were 0.85 and 0.86 respectively, reflecting a good balance between precision and recall.

Random Forest (RF) had the same accuracy as Logistic Regression, at 85%. This model is better at predicting patients without heart disease, with a precision of 0.86 and recall of 0.83. For class 1, precision was 0.85, slightly lower than Logistic Regression, but recall was higher

at 0.88, indicating that this model was better at detecting patients with liver disease. F1-score remained consistent, with values of 0.84 for class 0 and 0.86 for class 1.

K-Nearest Neighbors (KNN) has the best performance among all models with the highest accuracy of 92%. Precision for class 0 was 0.90 and for class 1 was 0.94, indicating that the prediction of patients with and without heart disease was highly accurate. Recall was also very high, with values of 0.93 for class 0 and 0.91 for class 1. F1 score was consistently high, at 0.92 for both classes, reflecting an excellent balance between precision and recall.

Based on the G-Mean calculation for the various classification models, K-Nearest Neighbors (KNN) has the highest score of 0.92. This indicates that KNN is the best model in maintaining a balance between detecting patients with and without heart disease. The Naive Bayes and Support Vector Classifier (SVC) models both have a G-Mean of 0.87, indicating a fairly balanced performance between the two classes. Random Forest and Logistic Regression have a G-Mean of 0.85 each, which means they are also quite good at handling class imbalance, although slightly lower than the other models.

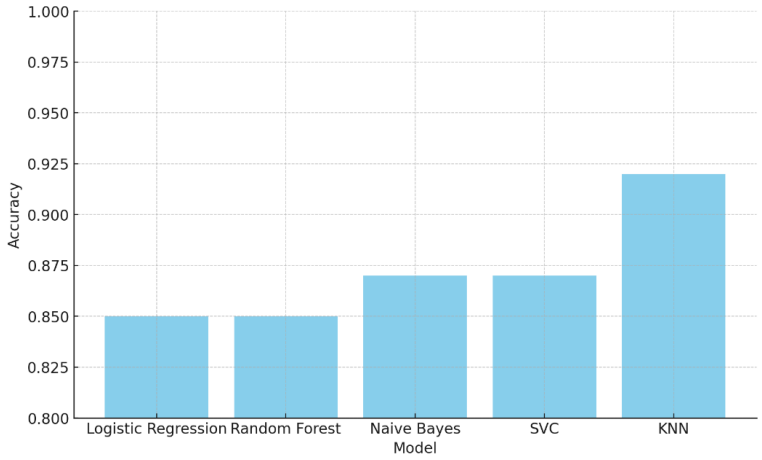

Figure 4. five classification models for accuracy

After comparing five classification models in predicting heart disease, K-Nearest Neighbors (KNN) proved to be the best performing model, having the highest accuracy of 92% and f1-score of 0.92 for both classes. Support Vector Classifier (SVC) and Naive Bayes (NB) also gave excellent results with 87% accuracy, showing a strong ability to maintain a balance between precision and recall. Random Forest and Logistic Regression had similar accuracy of 85%, but Random Forest excelled in recall for class 1, while Logistic Regression was more accurate in precision for class 1.

Although all models gave good results, K-Nearest Neighbors (KNN) was the overall best choice based on accuracy, precision, recall, f1-score and G-Mean metrics, making it the most reliable model to use for heart disease prediction with this dataset.

Overall, the performance of these models shows that no single model is significantly better than the others. However, KNN and SVC tend to be more reliable in this case, followed by Logistic Regression. These results can serve as a basis for choosing the most suitable model based on dataset complexity, speed, and ease of interpretation.

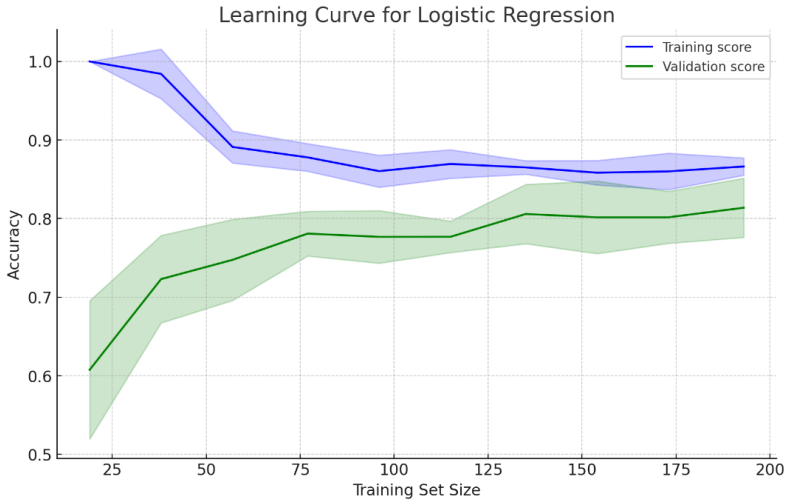

Figure 5. Learning Curve for Logistic Regression

The learning curve illustrates the learning process of the Logistic Regression model as the size of the training data increases. Initially, the training values (blue line) show very high accuracy when the training data size is small. This means that the model is able to remember the training data well, but it can also be an indication of overfitting as the model focuses too much on certain patterns in the small training data and cannot be generalised to other data.

However, as the size of the training data increases, the training score will decrease slightly. This decrease is normal and expected as the model starts to learn more general patterns that are not just specific to the small training data. This helps to reduce the risk of overfitting. At larger points, the training score reaches a stable value, which indicates that the model can recognise general patterns without focusing too much on specific training data.

On the other hand, the validation score (green line), which shows the model's performance on the test data (data not seen by the model during training), is initially lower than the training score. This indicates that the model cannot generalise well when the training data is small. However, as the size of the training data increases, the validation score gradually increases. This increase indicates that the model becomes better at generalising and is able to predict new data more accurately.

At some point, the two lines (training and validation scores) start to approach each other, which indicates that the model does not suffer from serious overfitting or underfitting. The model is able to maintain good performance on both training and test data, which means that

the model learns effectively and efficiently. The trend also shows that as the training data continues to grow, the model's performance tends to stabilise and become consistent.

Overall, this learning curve shows that Logistic Regression is able to handle the dataset well. The model continues to learn effectively as the data grows, achieving a balance between training accuracy and validation accuracy. This reflects the model's potential to provide better results with larger amounts of data, without showing significant signs of overfitting or underfitting.

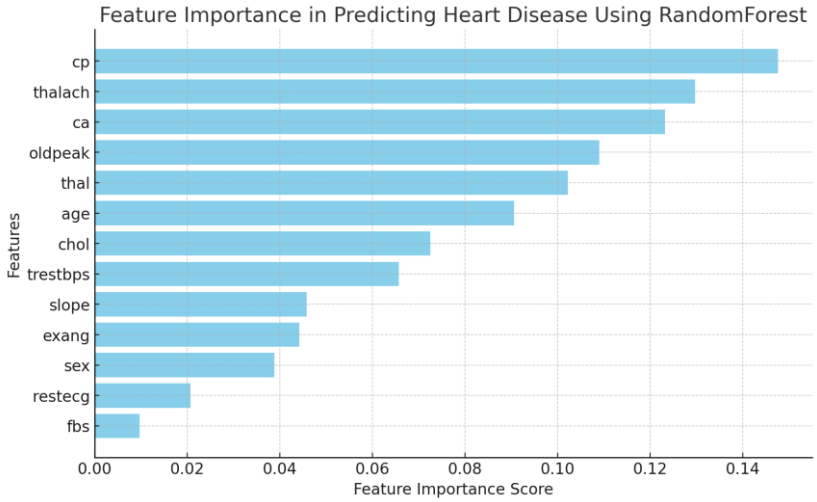

Figure 6. Feature Importance

Figure 6 shows the importance value of each feature in predicting heart disease using the Random Forest model. The most influential feature is cp (type of chest pain), with an importance value close to 0.15, which indicates that the type of chest pain of the patient is very influential in the diagnosis of heart disease. The feature thalach (maximum heart rate) also has a significant role, followed by ca (number of large blood vessels), oldpeak (ST depression), and thal (thalassaemia status).

These features, which have higher importance values, provide more relevant information in heart disease risk prediction. While other features such as gender, fbs (fasting blood sugar), and restecg (resting ECG results) have a smaller role, they still contribute to strengthening the overall prediction results.

Overall, the top features provide important information about the physical characteristics and health conditions of patients, which can be used to improve clinical diagnosis related to heart disease, as well as guide further decision-making for medical interventions.

5. DISCUSSION

A comparative analysis of five classification models-Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN)-revealed some important findings regarding their performance in predicting heart disease. Of all the models tested, K-Nearest Neighbors (KNN) showed the most superior performance with the highest accuracy of 92% as well as an excellent balance of precision, recall, and F1-score in both classes (class 0 and class 1). This indicates that KNN is the most reliable model in identifying patients with and without heart disease. The G-Mean score achieved by KNN, which is 0.92, further reinforces its ability to handle class imbalance well, making it a strong choice in this medical classification.

Naive Bayes (NB) and Support Vector Classifier (SVC) also performed very well with an accuracy of 87%. Both models have a high recall for class 0 of 0.90 and recall for class 1 of 0.84, indicating that both models have a good ability to classify patients both with and without heart disease. The F1-score of 0.87 in both models shows a good balance between precision and recall. The G-Mean of 0.87 also confirms that both models are strong enough to handle unbalanced datasets.

Meanwhile, Random Forest (RF) and Logistic Regression (LR) have the same accuracy of 85%, but with different characteristics. Random Forest excels in detecting patients with heart disease, with a recall value for class 1 of 0.88, making it a good choice for applications that require sensitive disease detection. On the other hand, Logistic Regression has a higher precision value in class 1 of 0.87, which means that this model is more accurate in predicting patients who actually have heart disease without generating too many false positives. The G-Mean of 0.85 for both models shows that they have sufficient ability to handle this classification, although their performance is slightly below that of KNN, NB, and SVC.

The results of the learning curve analysis for the Logistic Regression model indicate that the model learns effectively as the training data size increases. In the initial phase, the training score is high, which may indicate overfitting at small dataset sizes, where the model focuses more on the specific patterns of the training data. However, as the amount of data increases, the training score decreases slightly and stabilises, indicating that the model starts to recognise more representative general patterns. At the same time, the validation score increases gradually, indicating that the model is increasingly able to generalise to new data that it has not seen before. The convergence between training score and validation score shows that the model does not suffer from significant overfitting or underfitting, making Logistic Regression a reliable model for larger datasets.

From the perspective of feature importance using Random Forest, it was found that the most influential feature in predicting heart disease was chest pain type (cp), followed by thalach (maximum heart rate), ca (number of large blood vessels), oldpeak (exercise-induced ST depression), and thal (thalassemia status). These features play an important role in the diagnosis of heart disease, which can provide important insights for clinicians to determine appropriate medical intervention measures. Understanding the importance of these features can help improve the accuracy of diagnosis and strengthen decision-making in the clinical context.

Overall, the results of this study show that K-Nearest Neighbors (KNN) is the best model for heart disease classification in the context of this dataset. Support Vector Classifier (SVC) and Naive Bayes (NB) also performed very well, and are worth considering for further clinical applications. Random Forest and Logistic Regression although slightly inferior in some metrics, still provided competitive results, especially in terms of recall and precision specific to classification needs. The importance of clinical features such as chest pain, maximum heart rate, and ST depression, highlights the importance of these variables in heart disease risk prediction, which has direct implications in medical practice and research in the healthcare field.

6. CONCLUSIONS AND RECOMMENDATIONS

Based on the comparison of five classification models-Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN)-it is concluded that K-Nearest Neighbors (KNN) is the most superior model with the highest accuracy of 92%. KNN showed an excellent balance between precision, recall, F1-score, and G-Mean, making it the most reliable model in handling class imbalance in the heart disease prediction dataset. Support Vector Classifier (SVC) and Naive Bayes (NB) also performed well with 87% accuracy, showing a solid balance between precision and recall, making them strong alternatives in heart disease classification applications. Meanwhile, Random Forest (RF) and Logistic Regression (LR), despite having a slightly lower accuracy (85%), still show strong performance, where Random Forest is superior in detecting patients with heart disease (class 1), and Logistic Regression is more accurate in predicting patients who actually have heart disease.

For recommendations, firstly, K-Nearest Neighbors (KNN) is highly recommended for use in clinical prediction applications related to heart disease, especially in situations that require a balance between false positives and false negatives. Secondly, evaluation and monitoring of model performance needs to be done regularly to ensure the relevance and

accuracy of predictions, especially when there are changes in the patient population data. Thirdly, features such as chest pain type (cp), maximum heart rate (thalach), and ST depression (oldpeak), which have been shown to play an important role in heart disease prediction, should be of primary concern in clinical decision-making and further development of the prediction model. Fourth, further development through hyperparameter optimisation or combining multiple models (ensemble) can be done to improve prediction performance, especially on larger or more complex datasets. Finally, before widespread implementation of the model, additional validation using larger and more varied patient data is essential to ensure that the model is reliable in real clinical scenarios.

By following these recommendations, it is hoped that machine learning-based predictive applications in the field of heart disease can be more accurate and effective, and contribute to improving the quality of diagnosis and patient care in the future.

ADVANCED RESEARCH

In the context of further development, there are several areas that can be explored to improve the effectiveness of heart disease prediction using machine learning models. One approach is ensemble methods, such as bagging and boosting, which can improve model accuracy and stability. An ensemble model combines predictions from multiple base models to produce more robust and accurate predictions. Methods such as Gradient Boosting Machines (GBM), or XGBoost can be a good choice to strengthen classification performance by minimising bias and variance.

In addition, hyperparameter optimisation is an important step in honing model performance. Techniques such as Grid Search or Random Search can be used to find the best combination of model parameters that give optimal results. In further research, Bayesian Optimisation or Tree-structured Parzen Estimators (TPE) can also be used to find the optimal solution in less time and more efficiently than traditional search methods.

Another area that can be optimised is data processing through more in-depth feature engineering techniques. This process involves further exploration of existing features or even the creation of new features from the available data. Creating new features based on variable combinations or handling non-linear variables can have a significant impact on model performance. In addition, further research could explore dimensionality reduction techniques, such as Principal Component Analysis (PCA) or t-SNE, to reduce the complexity of the data without losing important information, so that the model can process the data more efficiently and effectively.

ACKNOWLEDGMENT

We would like to express our deepest gratitude to the colleagues who provided valuable advice during the process of writing this research. Your help, input and support have been an important part of the completion of this work. We would also like to express our appreciation to the colleagues who took the time to discuss and provide ideas that enriched this research.

In addition, we would like to express our gratitude to those who have provided financial support in the form of research grants that enabled this research to run smoothly, namely Universitas Informatika dan Bisnis Indonesia. This financial support is very meaningful in completing various stages of research and data analysis. Without the contribution and support from all parties, this research would not have achieved the expected results.

REFERENCES

Alarsan, F. I., & Younes, M. (2019). Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. *Journal of Big Data*, *6*(1), 81. https://doi.org/10.1186/s40537-019-0244-x

Ali, L., Niamat, A., Khan, J. A., Golilarz, N. A., Xingzhong, X., Noor, A., Nour, R., & Bukhari, S. A. C. (2019). An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure. *IEEE Access*, *7*, 54007–54014. IEEE Access. https://doi.org/10.1109/ACCESS.2019.2909969

Anjum, N., Siddiqua, C. U., Haider, M., Ferdus, Z., Raju, M. A. H., Imam, T., & Rahman, M. R. (2024). Improving Cardiovascular Disease Prediction through Comparative Analysis of Machine Learning Models. *Journal of Computer Science and Technology Studies*, *6*(2), Article 2. https://doi.org/10.32996/jcsts.2024.6.2.7

Cardiovascular diseases. (n.d.). Retrieved 29 September 2024, from https://www.who.int/health-topics/cardiovascular-diseases

Dutta, A., Batabyal, T., Basu, M., & Acton, S. T. (2020). An efficient convolutional neural network for coronary heart disease prediction. *Expert Systems with Applications*, *159*, 113408. https://doi.org/10.1016/j.eswa.2020.113408

Gárate-Escamila, A. K., Hajjam El Hassani, A., & Andrès, E. (2020). Classification models for heart disease prediction using feature selection and PCA. *Informatics in Medicine Unlocked*, *19*, 100330. https://doi.org/10.1016/j.imu.2020.100330

Gokulnath, C. B., & Shantharajah, S. P. (2019). An optimized feature selection based on genetic approach and support vector machine for heart disease. *Cluster Computing*, *22*(6), 14777–14787. https://doi.org/10.1007/s10586-018-2416-4

Ismail, A., Abdelrazek, S., & El-henawy, I. (2020). BIG DATA ANALYTICS IN HEART DISEASES PREDICTION. *Journal of Theoretical and Applied Information Technology*, *98*, 11.

Le, H. M., Tran, T. D., & Tran, L. V. (2018). AUTOMATIC HEART DISEASE PREDICTION USING FEATURE SELECTION AND DATA MINING TECHNIQUE. *Journal of Computer Science and Cybernetics*, *34*(1), Article 1. https://doi.org/10.15625/1813-9663/34/1/12665

Rahman, B., Hendric Spits Warnars, H. L., Subirosa Sabarguna, B., & Budiharto, W. (2021). Heart Disease Classification Model Using K-Nearest Neighbor Algorithm. *2021 Sixth International Conference on Informatics and Computing (ICIC)*, 1–4. https://doi.org/10.1109/ICIC54025.2021.9632918

Recursion Enhanced Random Forest With an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform | IEEE Journals & Magazine | IEEE Xplore. (n.d.). Retrieved 22 November 2023, from https://ieeexplore.ieee.org/abstract/document/9037283

Repaka, A. N., Ravikanti, S. D., & Franklin, R. G. (2019). Design And Implementing Heart Disease Prediction Using Naives Bayesian. *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 292–297. https://doi.org/10.1109/ICOEI.2019.8862604

Sekar, J., Aruchamy, P., Sulaima Lebbe Abdul, H., Mohammed, A. S., & Khamuruddeen, S. (2022). An efficient clinical support system for heart disease prediction using TANFIS classifier. *Computational Intelligence*, *38*(2), 610–640. https://doi.org/10.1111/coin.12487

Sireesha, M. (2020). Classification Model for Prediction of Heart Disease using Correlation Coefficient Technique. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*, 2116–2123. https://doi.org/10.30534/ijatcse/2020/185922020