Predicting Mortality in Heart Failure Patients: A Comparative Study of Decision Tree C4.5 and Naïve Bayes Algorithms

Didik Dwi Prasetya
Department of Electrical Engineering and Informatics
State University of Malang
Malang, Indonesia
didikdwi@um.ac.id

Muhammad Busthomi Arviansyah
Department of Electrical Engineering and Informatics
State University of Malang
Malang, Indonesia
Muhammad.busthomi.1805356@students.um.ac.id

Muhammad Taufiq Hidayat
Department of Electrical Engineering and Informatics
State University of Malang
Malang, Indonesia
muhammad.taufiq.2105356@students.um.ac.id

Heni Vidia Sari
Department of Electrical Engineering and Informatics
State University of Malang
Malang, Indonesia
heni.vidia.ft@um.ac.id

Azlan Mohd Zain
Faculty of Computing
Universiti Teknologi Malaysia
Johor Bahru, Malaysia
azlanmz@utm.my

2024 Ninth International Conference on Informatics and Computing (ICIC) | 979-8-3315-1760-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIC64337.2024.10956562

Abstract*—The heart is one of many vital organs of humans that is used to pump blood to the entire part of the body. One of many diseases that attack the heart is heart failure. Heart failure arises from either structural or functional abnormalities that hinder the ventricle's capability to effectively pump blood throughout the body. Based on the data released by the Health Research Department, the prevalence of heart failure in Indonesia is 1.5%, which means 1 million Indonesians or more have heart problems. The mortality rate of heart failure patients can be reduced by data mining classification. The classification process uses decision tree C4.5 and the naïve Bayes algorithm. From 299 data of heart failure patients that were classified using a decision tree, the C4.5 algorithm got an 83.26% accuracy rate, and the naïve Bayes algorithm got a 79.26% accuracy rate. From the classification process that has been done, it can be seen that the decision tree C4.5 algorithm has better performance than the naïve Bayes algorithm in the classification process of heart failure patient data.

Keywords: heart failure, classification, decision tree C4.5, naïve Bayes

I. INTRODUCTION

The heart is one of the most essential organs, and it pumps blood to the whole body. One of the various diseases of the heart is heart failure. Heart failure is the final problem that can occur in the heart. Heart failure is a multifaceted syndrome resulting from either structural or functional abnormalities that impair the ventricle's ability to effectively pump blood [1]. This condition occurs when the heart is unable to circulate sufficient blood to fulfill the body's demands for nutrients and oxygen, leading to ventricular dilation [2].

Heart failure is the leading medical problem, with 23 million prevalence all over the world [3]. Based on the data released by health researchers in Indonesia in 2018, the prevalence of heart disease in Indonesia is 1.5% [4]. That means more than 4 million Indonesians have problems with their heart. The mortality of heart failure patients can be decreased by seeking the indications that cause the heart failure patient to pass away. With the current technological development, the mortality caused by heart failure can be reduced using the data mining technique.

The data mining technique used in this study is classification, and the data is the Heart Failure Clinical Records Dataset from the UCI Machine Learning Repository. This study is comparing the classification algorithms C4.5 and Naive Bayes. Based on the same comparative study using the coronary heart dataset, the accuracy rate of the C4.5 and Naive Bayes algorithm is equal [5]. The other comparison study between those two algorithms using the diabetics dataset found that the accuracy rate obtained by the Naive Bayes algorithm is higher than the accuracy rate of the C4.5 algorithm [6]. The former studies show a bias about the effectiveness of the two algorithms. This study compared the C4.5 and Naive Bayes algorithms using heart failure patient data.

Heart failure is defined as the condition where the heart is not able to pump enough blood to the body's network. The heart problem can be caused by systolic and diastolic functions, arrhythmia, or a mismatch between the preload and afterload [2][7]. The part of the heart can classify heart failure, left or right. Heart failure also can be classified into acute heart failure, chronic decompensated heart failure, and chronic heart failure [3][8].

Many things can cause heart failure. Epidemiologically, it is essential to know the cause of the heart failure. The typical symptoms of heart failure are: hard to breathe, fatigue, and leg edema. At the same time, the usual signs of heart failure are tachycardia, tachypnea, Ronchi voice, pleural effusion, jugular vein distention, peripheral edema, and hepatomegaly [2][9]. By knowing the causes and symptoms of heart failure, early treatment would be beneficial in the heart's healing process.Related Works

Data mining is a process used to gain information from the data that is useful for making a decision. Data mining encompasses various branches, including Database Systems, Data Warehousing, Statistics, Machine Learning, Information Retrieval, and Advanced Computing Techniques. Data mining is a process of gaining information from big data using Machine learning, statistics, and Database Management System algorithms and techniques [10]. Data mining has many functions, but the main tasks are predictive and descriptive functions [11]. The data mining technique used in this research is classification

Classification is a technique used to know or estimate the class of the object based on the characteristic similarity of the data. The attribute represents the characteristic of the object where the object has various values. Not all of the attributes affect the classification process. Classification algorithms in machine learning include Naïve Bayes, Support Vector Machines, Decision Trees, Fuzzy Logic, and Artificial Neural Networks [11]. These algorithms are commonly used in technology-assisted classification tasks. Previous studies have demonstrated the application of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) algorithms in heart disease prediction, showcasing their ability to classify cardiovascular risk factors with varying levels of accuracy [12][13].

The C4.5 algorithm is extensively used for constructing decision trees, which are renowned for their effectiveness in classification and prediction tasks. C4.5 is appreciated for its straightforwardness and ease of implementation, making it one of the most user-friendly classification algorithms available [14]. One of the most straightforward and frequently employed classification methods is the Naïve Bayes classifier. It functions based on a probabilistic approach, utilizing Bayes' theorem while making a strong (naive) assumption of independence among features [15]. Despite its simplicity, Naïve Bayes proves to be highly effective for a variety of classification tasks. In this study, a comparison is made between the Naïve Bayes algorithm and the C4.5 algorithm to identify which provides greater accuracy in detecting heart disease.

II. Methods

A. Research Model

The research flow in this study is illustrated in Figure 1. From Figure 1. the first step of this research is the dataset step. In this step, pre-processing data is carried out. The next step is feature selection. In this step, the attributes of the dataset are selected to find the most relevant attributes to make the classification process more efficient. The next step of the research is the classification itself. C4.5 and the Naive Bayes algorithm were used for the classification process. The last step of the research is validation and evaluation.

The validation process is used to validate the classification process. After the validation step, the next step is evaluation. This step is used to evaluate which algorithm is better to classify the data of heart failure patients. The evaluation step calculates the accuracy rate of each algorithm. The accuracy rate can be obtained using the confusion matrix and accuracy formula, as depicted in Table 1.

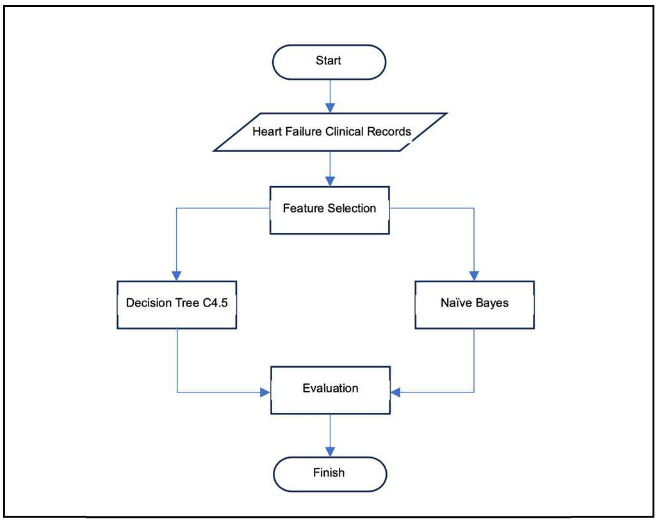

Fig. 1. Research flow

TABLE I. Confusion Matrix

Correct Clustering	Clustered as	
	+	-
+	True Positives	False Positives
-	False Negatives	True Negatives

B. Dataset

The dataset used in this study is the Heart Failure Clinical Records Dataset. The dataset description is provided in Table 2. This public dataset was obtained from the UCI Machine Learning Repository website. The dataset has 299 data on heart failure patients, with the indications as the dataset's attribute. In this dataset, the "death_events" attribute acts as the target class in which the '1' value indicates that the patient passed away during the following period of heart failure, and the '0' value indicates that the patient is still alive during the subsequent period.

TABLE II. Description of Dataset

No	Attributes	Description
1.	Age	Age
2.	Anaemia	Anaemia
3.	High_blood_pressure	High blood pressure
4.	Creatinine_phospokinase	Level of creatinine phosphokinase
5.	Diabetes	Diabetes
6.	Ejection_fraction	Blood percentage when the heart is pumping
7.	Sex	Gender
8.	Platelets	Platelets
9.	Serum_creatinine	Creatinine
10.	Serum_sodium	Sodium
11.	Smoking	Smoking
12.	Time	Time
13.	Death_events	Death status

C. Pre-processing

Data pre-processing is a crucial step. In data mining, this process can affect the result of the data itself. The unprocessed data usually contains some errors or unexpected values when the data is being used. The errors in the raw data are caused by the inconsistency of the data itself. The incomplete data exists because of the missing values in the attribute of the raw data [12].

D. Feature Selection

Feature selection has been effective and efficiently proven in its use at the pre-processing stage of data mining and machine learning [16]. The objective of feature selection is to identify the subset of most pertinent features to be utilized in constructing the model [17]. Feature Selection can make the computing process more efficient. Besides that, this process can improve the performance of an algorithm. One of the types of feature selection is the filter method. This method is done before the classification process runs. This method is independent of the learning algorithm employed in the classification process. By utilizing filter methods, the feature selection process is performed once, and the resulting selected features can subsequently serve as input for various classifiers [18] [19]. The filter method used in this study is correlation attribute evaluation.

III. Results and Discussion

Classification is a technique used to know or estimate the class of the object based on the characteristic similarity of the data. The attribute represents the characteristic of the object where the object has various values. Not all of the attributes have an effect on the classification process. This study involves two scenarios: classification with feature selection and classification without feature selection.

From the feature selection process, which has been done before, five attributes are used as input for the classification process. The attributes considered in this study included time, serum creatinine, ejection fraction, age, and serum sodium. The classification process was carried out twice. The first one was done with the feature selection process, and the second one was done without the feature selection process.

A. Decision Tree C4.5

The decision tree is the most known algorithm that is used to make a decision. This algorithm is one of the most popular classification algorithms to use [10]. The decision tree is a predictive decision model that uses a tree structure [16]. The decision tree C4.5 model has more advantages over the ID3 and CART models because of its ability not to limit the branch in binary form.

The Decision Tree C4.5 algorithm performs noticeably better than another method, as evidenced by the results presented in Table 3, which clearly highlight its effectiveness in accurately classifying the dataset under consideration.

TABLE III. The result of C4.5

Algorithm	*Feature Selection*	Accuracy
C4.5	✓	83.26%
C4.5	✕	79.93%

B. Naive Bayes

Naive Bayes is a classification algorithm within the realm of data mining that leverages Bayesian theory for its predictive capabilities [18]. The Bayes prediction theorem serves as a foundational statistical method for recognizing patterns. This algorithm operates under the simplifying assumption that, given a specific output value, the attributes' values are conditionally independent of one another [20].

Although Naive Bayes demonstrates results that fall below those of Decision Tree C4.5, as illustrated in Table 4, its performance remains acceptable and still provides significant contributions within the context of the analysis conducted.

TABLE IV. The result of Naive Bayes

Algorithm	*Feature Selection*	Accuracy
Naïve Bayes	✓	79.26%
Naive Bayes	✕	76.58%

C. Classification Analysis

From Table 3, the C4.5 classification process with feature selection gave a better accuracy rate than the C4.5 classification without feature selection. The accuracy rate of the classification process using the feature selection is 83.26%, and the accuracy rate of the non-feature selection classification is 79.93%.

The C4.5 excels due to its ability to handle both categorical and continuous data, efficiently perform information gain ratio-based splits, manage missing values, and utilize pruning techniques to enhance model generalization, making it suitable for various classification tasks. Additionally, C4.5's robust handling of noisy data and its ability to generate easily interpretable decision trees contribute to its widespread applicability in real-world scenarios. The algorithm's capacity to produce compact models while maintaining high accuracy allows for effective decision-making in complex datasets.

From Table 4, the Naive Bayes classification process with feature selection gave a better accuracy rate than the Naive Bayes classification without feature selection. The classification process utilizing feature selection achieved an accuracy rate of 79.26%, while the accuracy rate for the classification without feature selection was recorded at 76.58%.

Feature selection enhances accuracy by reducing the dimensionality of the dataset, which minimizes the risk of overfitting. By identifying and retaining only the most relevant features, it eliminates noise and irrelevant information, allowing machine learning algorithms to focus on the most informative data points. This simplification can lead to improved model performance, as the algorithms can learn more effectively from a cleaner, more relevant dataset. Furthermore, feature selection can improve computational efficiency and reduce training time, enabling quicker model convergence and better generalization to unseen data.

The results indicate that incorporating feature selection improves the accuracy of both the C4.5 and Naive Bayes classification processes. This condition aligns with previous research findings that demonstrate the superiority of C4.5 in classification tasks [5][6]. "Nonetheless, Naïve Bayes also holds significant potential and remains a viable option worthy of consideration in various classification scenarios [6]. Thus, the data mining approach offers considerable potential for solving problems efficiently and accurately by uncovering hidden patterns, relationships, and trends within large datasets, enabling timely and data-driven decision-making [21].

Feature selection can introduce overhead due to the additional computational resources required to evaluate and identify the most relevant features from the dataset. This process often involves multiple iterations of calculating feature importance or applying selection algorithms, which can increase processing time and complexity, especially in large datasets. However, in the study, which involves a dataset of 299 numeric records, no significant overhead was observed. The relatively small size of the dataset likely mitigated any computational burden, allowing the feature selection process to run efficiently without negatively impacting the performance of the classification algorithms.

IV. Conclusion

This study explores the transformative potential of data mining techniques in addressing the pressing public health challenge of heart failure. By leveraging the decision tree C4.5 and naïve Bayes algorithms, the research sought to refine prediction and management strategies for this critical condition. The comparative analysis underscores the pivotal role of feature selection in enhancing predictive accuracy. Notably, regardless of feature selection, the decision tree C4.5 algorithm consistently outperformed the naïve Bayes approach in classifying heart failure patient data. These findings not only highlight the efficacy of data-driven methodologies in bolstering diagnostic precision but also suggest actionable pathways for optimizing treatment strategies. Ultimately, this research emphasizes the significance of innovative data mining techniques in mitigating the burden of heart failure, thus paving the way for improved patient care and outcomes.

Acknowledgment

The authors extend sincere gratitude to the Department of Electrical Engineering and Informatics, State University of Malang, for the invaluable support and encouragement throughout the research and publication process.

References

[1] S. A. Hunt, D. W. Baker, and M. H. Chin, "ACC/AHA guidelines for the evaluation and management of chronic heart failure in the adult: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines," *J. Am. Coll. Cardiol.*, vol. 38, no. 12, pp. 2101-2113, 2001.

[2] N. Moghaddam, N. Malhi, and M. Toma, "Impact of oral soluble guanylate cyclase stimulators in heart failure: A systematic review and meta-analysis of randomized controlled trials," *Am. Heart J.*, vol. 241, pp. 74-82, 2021.

[3] A. Groenewegen, F. H. Rutten, A. Mosterd, and A. Hoes, "Epidemiology of heart failure," *Eur. J. Heart Fail.*, vol. 22, no. 8, pp. 1342-1356, 2020.

[4] Kementerian Kesehatan RI, "Situasi Kesehatan Jantung," Pusat Data dan Informasi Kementerian Kesehatan RI, 2018.

[5] A. Husejinovic, "Credit card fraud detection using naive Bayesian and C4.5 decision tree classifiers," *Credit Card Fraud Detection Using Naive Bayesian and C4.5*, vol. 4, pp. 1-5, 2020.

[6] I. G. A. Suciningsih, M. A. Hidayat, and R. A. Hapsari, "Comparative analysis of naïve Bayes and decision tree C4.5 for caesarean section prediction," *J. Soft Computing Explor.*, vol. 2, no. 1, pp. 46-52, 2021.

[7] J. P. Moore et al., "Management of heart failure with arrhythmia in adults with congenital heart disease: JACC state-of-the-art review," *J. Am. Coll. Cardiol.*, vol. 80, no. 23, pp. 2224-2238, 2022.

[8] J. Njoroge and J. R. Teerlink, "Pathophysiology and therapeutic approaches to acute decompensated heart failure," *Circ. Res.*, vol. 128, no. 10, pp. 1468-1486, 2021.

[9] T. Smole et al., "A machine learning-based risk stratification model for ventricular tachycardia and heart failure in hypertrophic cardiomyopathy," *Comput. Biol. Med.*, vol. 135, 104648, 2021.

[10] M. K. Gupta and P. Chandra, "A comprehensive survey of data mining," *Int. J. Inf. Technol.*, vol. 12, no. 4, pp. 1243-1257, 2020.

[11] A. Dogan and D. Birant, "Machine learning and data mining in manufacturing," *Expert Syst. Appl.*, vol. 166, 114060, 2021.

[12] R. Ahmed, M. Bibi, and S. Syed, "Improving heart disease prediction accuracy using a hybrid machine learning approach: A comparative study of SVM and KNN algorithms," *Int. J. Computations, Inf. Manuf.*, vol. 3, no. 1, pp. 49-54, 2023.

[13] B. Duraisamy, R. Sunku, K. Selvaraj, V. V. R. Pilla, and M. Sanikala, "Heart disease prediction using support vector machine," *Multidisciplinary Sci. J.*, vol. 6, 2024.

[14] T. H. Sinaga et al., "Implementation of data mining using C4.5 algorithm on customer satisfaction in Tirta Lihou PDAM," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 3, no. 1, pp. 9-20, 2021.

[15] B. Ravinder et al., "Web data mining with organized contents using Naive Bayes algorithm," in *2024 2nd Int. Conf. Computer, Communication and Control (IC4)*, 2024, pp. 1-6.

[16] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2012.

[17] J. Li et al., "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, article 94, 2017.

[18] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157-1182, 2003.

[19] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection," *Int. J. Inf. Technol. Knowledge Manag.*, vol. 2, no. 2, pp. 271-277, 2010.

[20] D. D. Prasetya, A. P. Wibawa, and T. Hirashima, "The performance of text similarity algorithms," *Int. J. Adv. Intell. Informatics*, vol. 4, no. 1, pp. 63-69, 2018.

[21] D. D. Prasetya and T. Hirashima, "Associated patterns in open-ended concept maps within e-learning," *Knowl. Eng. Data Sci.*, vol. 5, no. 2, pp. 179-187, 2022.