

CHAPTER I INTRODUCTION

1.1 Background

Heart disease is a disease in which the heart stops functioning, heart disease has various causes such as blocked arteries, inflammation, infection, or birth defects [1]. According to a data from the World Health Organization (WHO), heart disease accounts for around 31% of total global deaths, with more than 17 million people dying each year from coronary heart disease. Heart disease is also a disease that is increasing very rapidly, where at least 15 out of every 1,000 people in Indonesia suffer from heart disease. According to research [2], heart disease is the leading cause of death in the world, so early detection and accurate classification will be very important for proper treatment.

In recent years, the usage of newly artificial intelligence (AI) technology in the healthcare sector has grown. One method widely to be used in health data analysis is machine learning, particularly in classification tasks. Heart disease classification aims to identify high-risk patients based on their recorded clinical risk factors. For the context of this study, the patient medical records used were sourced from the UCI Heart Disease dataset, which presents both text-based and numeric (tabular) data. This dataset records various essential patient parameters, ranging from demographics (age, gender), symptomatic complaints (type of chest pain), to physiological metrics (cholesterol, blood pressure, and blood sugar). One machine learning model that has proven robust for classifying such high-dimensional tabular data is Random Forest.

However, in practice, Random Forest often faces issues related to relevant feature selection and data imbalance, which can degrade model accuracy. Therefore, algorithm optimization and data preprocessing techniques such as Particle Swarm Optimization (PSO), feature selection using Boruta, and data balancing techniques such as SMOTE-Tomek are crucial for improving model performance.

PSO, firstly introduced by Kennedy and Eberhart in 1995, is a population-based optimization method inspired by the social behavior of many animals, such as fish and birds. In PSO, each individual (called a particle) in the population attempts to find the optimal solution based on the best position found by itself and the best position that found by other individuals in population. This process occurs through iterations that update the particle's position and velocity to achieve the best solution. PSO is used

for various applications, one of which is in hyperparameter optimization in models of machine learning, where it helps to find the best parameter configuration that can improve the performance and accuracy of any models, such as Random Forest, SVM, or artificial neural networks. Research [3] conducted in evaluating the effect of using PSO in coral health classification showed that model accuracy increased from 79.11% to 85.44% after optimization with PSO. In addition, computation time was also significantly reduced. PSO has been shown to improve the ability of many models to recognize patterns as well as speed up the computation process, which make it suitable for any large-scale or real-time applications.

The use of Boruta as an effective method for feature selection in selecting the features that most relevant in a classification model, especially when the dataset has many irrelevant features. This algorithm uses Random Forest as a basis for determining which features are important by trying to compare the predictive power of each existing feature with features that randomly generated. Boruta functions to identify and retain any significantly influence features on the model's results, as well as eliminating features that not significantly contribute to the model's decision. This will also helps improve model accuracy as well as strengthens the interpretation of the results, because only relevant features are retained in the analysis process. In research [4] it is explained that Boruta feature selection selects 6 of the most relevant features thereby improves the performance of the classification model in predicting heart disease. In other words, Boruta helps filter out important features so that the classification algorithm becomes more effective and accurate in predicting results.

And the usage of SMOTE-Tomek as a method to address any class imbalance problems in the dataset, where the amount of data for the minority class (for example, patients with heart disease) is much less than the majority class. SMOTE (Synthetic Minority Over-sampling Technique) will synthesize new data samples for minority class by interpolating between original samples. While Tomek Links are used to identify and remove conflicting data pairs between the majority and minority class. So the application of these two techniques combination, SMOTE-Tomek, can improve the generalization and quality of the classification model by addressing class imbalance. Research [5] shows that the use of SMOTE-Tomek can improve the performance and quality of the Naive Bayes algorithm, especially in increasing the ROC-AUC score, which means the model is better at distinguishing between classes. Balanced data produces a better model in detecting heart disease, although the increase

in accuracy is not always significant, but there are improvements in other metrics such as precision, recall, and f-measure.

The combination of these three techniques is expected to improve the quality, accuracy, and generalization of the Random Forest model in classifying heart disease. By optimizing the model using PSO, performing better feature selection with Boruta, and balancing the dataset using SMOTE-Tomek, the model is expected to provide more accurate and reliable results in the early detection of heart disease.

1.2 Problem Formulation

Based on the identified problems in the background, this study formulates several main questions to be answered, namely:

1. How does the PSO-Random Forest model perform with the Boruta feature selection implementation in heart disease classification?
2. How does the PSO-Random Forest model perform with the implementation of the SMOTE-Tomek data balancing technique in heart disease classification?
3. How does the PSO-Random Forest model perform with the implementation of Boruta feature selection and SMOTE-Tomek data balancing techniques in heart disease classification?

1.3 Research Objectives

This research aims to analyze and evaluate the performance of Random Forest models optimized by Particle Swarm Optimization by combining the Boruta feature selection algorithm and the SMOTE-Tomek data balancing technique using several different approaches. Based on the problems explained previously, the main objectives to be achieved in this research are as follows:

1. To determine the performance of the PSO-Random Forest model with the implementation of Boruta feature selection in improving the accuracy of heart disease classification.
2. To determine the performance of the PSO-Random Forest model with the implementation of the SMOTE-Tomek data balancing technique in overcoming the problem of data imbalance in the heart disease dataset.
3. To determine the performance of the PSO-Random Forest model with the implementation of Boruta feature selection and SMOTE-Tomek data balancing techniques, and to evaluate whether the combination of these techniques can improve the accuracy of the model in heart disease classification.

1.4 Problem Definition

In order for this research to have a clear focus and be able to be completed in a measurable manner, it is necessary to establish several problem limitations as follows:

1. This study only uses secondary datasets obtained from the Kaggle site, namely UCI Heart Disease Data, which consists of 920 data.
2. This study only uses the Random Forest algorithm optimized with PSO as the main classification model.
3. The Boruta feature selection technique and SMOTE-Tomek data balancing were used only for comparison of the results of the main model in heart disease classification.
4. The research focus is on the analysis of model evaluation using evaluation metrics like accuracy, precision, sensitivity, and F1-Score.

1.5 Benefits of Research

This research is expected to make a significant contribution, both theoretically and practically, particularly in the development of information technology and health sciences. The benefits to be gained from this research are as follows:

1. The results of this research are hopefully be beneficially to the development of science, particularly in the machine learning and its application in healthcare sector. The use of PSO optimization techniques, Boruta feature selection, and SMOTE-Tomek data balancing in heart disease classification is expected to be contributing to the development of more improved classification models.
2. This research is to be expected to provide real support to the world of health, by providing more efficient tools in detecting and diagnosing heart disease, so that it can help medical personnel in providing more precise and faster interventions.
3. This research also expected to be able to provide benefits for technology development by introducing a combination of techniques that can improve the quality of the Random Forest model, which may be applied to other classification problems.