HYPERPARAMETER TUNING OPTIMIZATION USING GRIDSEARCH AND RANDOMSEARCH IN SUPPORT VECTOR MACHINE (SVM) METHOD FOR THE CLASSIFICATION OF HEART DISEASE PATIENT DATA

Thesis

By

Dinda Meilani Aditya Wati
2117031060

DEPARTMENT OF MATHEMATICS
FACULTY OF MATHEMATICS AND NATURAL SCIENCE
UNIVERSITY OF LAMPUNG
2025

ABSTRACT

HYPERPARAMETER TUNING OPTIMIZATION USING GRIDSEARCH AND RANDOMSEARCH IN SUPPORTVECTOR MACHINE (SVM) METHOD FOR THE CLASSIFICATION OF HEART DISEASE PATIENT DATA

By

DINDA MEILANI ADITYA WATI

Heart disease is one of the most deadly Non-Communicable Diseases (NCD) in the world and is the leading cause of death in Indonesia. Efforts to early detection of heart disease are very important to increase the effectiveness of treatment and reduce the risk of death. In this study, the Support Vector Machine (SVM) method is used to classify data of heart disease patients. This study aims to optimize the performance of SVM models through hyperparameter tuning techniques using GridSearch and RandomSearch, as well as the application of data balancing methods, namely Random Oversampling. The data used is secondary data obtained from the Kaggle site, consisting of 918 patient data with 11 independent variables and 1 target variable. The research process includes data preprocessing stages (cleaning, feature selection, balancing, scaling), modeling with SVM, and model performance evaluation using accuracy, precision, recall, and f1-score metrics. The results show that hyperparameter tuning optimization using RandomSearch with Random Oversampling technique produces the best performance with an accuracy of 89.21%, compared to GridSearch and the default model which has an accuracy of 85.29%. Thus, RandomSearch is proven to be more effective in improving the performance of SVM classification models to detect heart disease.

Keywords: Support Vector Machine, Hyperparameter Tuning, GridSearch, RandomSearch, Heart Disease, Random Oversampling.

ABSTRAK

OPTIMASI HYPERPARAMETER TUNING MENGGUNAKAN GRIDSEARCH DAN RANDOMSEARCH PADA METODE SUPPORT VECTOR MACHINE (SVM) UNTUK KLASIFIKASI DATA PASIEN PENYAKIT JANTUNG

Oleh

DINDA MEILANI ADITYA WATI

Penyakit jantung merupakan salah satu penyakit tidak menular yang paling mematikan di dunia dan menjadi penyebab utama kematian di Indonesia. Upaya deteksi dini terhadap penyakit jantung sangat penting untuk meningkatkan efektivitas penanganan dan mengurangi risiko kematian. Dalam penelitian ini, digunakan metode Support Vector Machine (SVM) untuk melakukan klasifikasi data pasien penyakit jantung. Penelitian ini bertujuan untuk mengoptimalkan kinerja model SVM melalui teknik hyperparameter tuning menggunakan GridSearch dan RandomSearch, serta penerapan metode balancing data yaitu Random Oversampling. Data yang digunakan merupakan data sekunder yang diperoleh dari situs Kaggle, terdiri dari 918 data pasien dengan 11 variabel independen dan 1 variabel target. Proses penelitian meliputi tahap preprocessing data (cleaning, feature selection, balancing, scaling), pemodelan dengan SVM, serta evaluasi performa model menggunakan metrik akurasi, presisi, recall, dan f1-score. Hasil penelitian menunjukkan bahwa optimasi hyperparameter tuning menggunakan RandomSearch dengan teknik Random Oversampling menghasilkan performa terbaik dengan akurasi sebesar 89,21%, dibandingkan dengan GridSearch maupun model default yang memiliki akurasi sebesar 85,29%. Dengan demikian, RandomSearch terbukti lebih efektif dalam meningkatkan performa model klasifikasi SVM untuk mendeteksi penyakit jantung.

Kata kunci: Support Vector Machine, Hyperparameter Tuning, GridSearch, RandomSearch, Penyakit Jantung, Random Oversampling

HYPERPARAMETER TUNING OPTIMIZATION USING GRIDSEARCH AND RANDOMSEARCH IN SUPPORT VECTOR MACHINE (SVM) METHOD FOR THE CLASSIFICATION OF HEART DISEASE PATIENT DATA

By

DINDA MEILANI ADITYA WATI

Thesis

Submitted as a Partial Fulfilment of the Requirement for the Degree of
BACHELOR OF MATHEMATICS

In the

Department of Mathematics
Faculty of Mathematics and Natural Science

FACULTY OF MATHEMATICS AND NATURAL SCIENCE
UNIVERSITY OF LAMPUNG
BANDAR LAMPUNG
2025

Thesis Title	: **HYPERPARAMETER TUNING OPTIMIZATION USING GRIDSEARCH AND RANDOMSEARCH IN SUPPORT VECTOR MACHINE (SVM) METHOD FOR THE CLASSIFICATION OF HEART DISEASE PATIENT DATA**
Student Name	: Dinda Meilani Aditya Wati
Student ID Number	: **2117031060**
Program	: **Mathematics**
Faculty	: **Mathematics and Natural Sciences**

APPROVED

1. Supervisory Committee

Dr. Khoirin Nisa, S.Si., M.Si.
NIP 197407262000032001

Misgiyati, S.Pd., M.Si.
NIP 198509282023212032

2. Chair of the Department of Mathematics

Dr.Aang Nuryaman, S.Si.,M.Si.
NIP.197403162005011001

APPROVED

1. Examination Team

Chairperson : **Dr. Khoirin Nisa, S.Si., M.Si.** .....

Secretary : **Misgiyati, S.Pd., M.Si.** .....

Examiner
not Supervisor : **Widiarti, S.Si., M.Si.** .....

2. Dean of the Faculty of Mathematics and Natural Sciences

Dr. Eng. Heri Satria, S.Si., M.Si.
NIP. 197110012005011002

Date of Thesis Examination: **19 August 2025**

STUDENT THESIS STATEMENT

The undersigned :

Name	:	**Dinda Meilani Aditya Wati**
Student ID Number	:	**2117031060**
Department	:	**Mathematics**
Thesis Title	:	**HYPERPARAMETER TUNING OPTIMIZATION USING GRIDSEARCH AND RANDOMSEARCH IN SUPPORT VECTOR MACHINE (SVM) METHOD FOR THE CLASSIFICATION OF HEART DISEASE PATIENT DATA**

I hereby declare that this thesis is my own work. If it is later proven that this thesis is a copy or was written by someone else, I am willing to accept sansctions in accordance with applicable academic regulations.

Bandar Lampung, 19 August 2025

Author

Dinda Meilani Aditya Wati

BIOGRAPHY

The author's full name is Dinda Meilani Aditya Wati, born on May 18, 2003, in Bandar Lampung, Lampung. The author is the third of three children of Mr. Wagino and Mrs. Musinah.

The author attended kindergarten at Gula Putih Mataram Kindergarten from 2007 to 2009, elementary school at Gula Putih Mataram Elementary School from 2009 to 2015, and junior high school at Gula Putih Mataram Junior High School from 2015 to 2018. Then, from 2018 to 2021, the author attended high school at SMAS Sugar Group Bandar Mataram, Central Lampung, Lampung.

In 2021, the author continued her education in collage and enrolled as an undergraduate student in the Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Lampung through the Join Selection for State Universities (SBMPTN). During her time as a student, the author was active in BEM FMIPA as an expert staff member for Women's Empowerment (PW).

From December 2023 to February 2024, the author conducted an internship at PT Great Giant Pineapple. Subsequently, the author carried out a Field Study Program (KKN) from June to August 2024 in Karya Makmur Village, Labuhan Maringgai, East Lampung.

INSPIRATIONAL QUOTE

"And that man only reaps what he has sown”

(Q.S. An-Najm: 39)

“No results without effort, no success without patience.”

“Life is not about competing with each other, dream for yourself.”

“So when you have finished (one task), continue working hard (on another). And put your hope in God alone.”

(Q.S. Al-Insyirah: 7–8)

“No dream is a failure; there are only dreams that are delayed. If you feel like you have failed to achieve your dream, don't worry. You can always create new dreams.”

DEDICATION

With gratitude and thanks to Allah SWT for His blessings and guidance, this thesis has been competed successfully and on time. With gratitude and happiness, I would like to express my gratitude to :

My Beloved Father and Mother

Thank you to my parents for all your sacrifies, motivation, prayers, blessings, and support throughout my life. Thank you for teaching me valuable lessons about the true meaning of life's journey so that I can become someone who is useful to many people.

Supervisors and Discussant

Thank you to the supervisors and discussant who have been extremely helpful, providing motivation, guidance, and valueable knowledge.

My friends

Thank you to all the kind people who have shared their experiences, enthusiasm, motivation, prayer, and unwavering support in all matters.

Beloved Alma Mater

University of Lampung

SANWACANA

Praise be to Allah, the author offers thanks and gratitude to Allah SWT for His abundant mercy and blessings, enabling the author to complete this thesis titled "Hyperparameter Tuning Optimization Using Gridsearch and Randomsearch in the Support Vector Machine (SVM) Method for the Classification of Heart Disease Patient Data" successfully, smoothly, and on time. May blessings and peace be upon the Prophet Muhammad SAW.

In the process of writing this thesis, many people have helped by providing guidance, support, direction, motivation, and advice so that this thesis could be completed. Therefore, the author would like to take this opportunity to express his gratitude to:

1. Mrs. Dr. Khoirin Nisa, S.Si., M.Si., as Supervisor 1, who has devoted a great deal of her time to providing guidance, direction, motivation, advice, and support to the author, enabling the completion of this thesis.
2. Mrs. Misgiyati, S.Pd., M.Si., as Supervisor II, who has provided guidance and advice to the author throughout the process of writing this thesis.
3. Mrs. Widiarti, S.Si., M.Sc., as the Discussant who always provided advice and criticism regarding the writing of this thesis.
4. Mr. Dr. Aang Nuryaman, S.Si., M.Si., as the Head of the Mathematics Department, Faculty of Mathematics and Natural Sciences, University of Lampung.
5. Mr. Dr. Ahmad Faisol, S.Si., M.Sc., as the Secretary of the Mathematic Department, Faculty of Mathematics and Natural Sciences.
6. Mrs. Dr. Dian Kurniasari, S.Si., M.Sc., as the academic advisor who has consistently guided the author throughout the higher education process.

7. All lecturers, staff, and employees of the Department of Mathematics, Facu of Mathematics and Natural Sciences, University of Lampung.
8. Mr. Dr. Eng. Heri Satria, S.Si., Dean of the Faculty of Mathematics and Natural Sciences, University of Lampung.
9. Dear family, father, mother, siblings, and extended family who have consistently provided support, motivation, and prayers to the author.
10. The author's friends, Dian, Imas Happy, Made Dinda, Miranda, and Civita, who have provided support and always accompanied the author.
11. My college friends Yulina, Ana, Dea, Cantika, Rani, and Lisa, who have provided encouragement and support.
12. My friends from the Mathematics Department, Class of 2021, who have been with me throughout my studies.

May this thesis be of benefit. The author realizes that this thesis is far from perfect, so the author welcomes constructive criticism and suggestions to make this thesis even better.

Bandar Lampung, August 19, 2025

Dinda Meilani Aditya Wati

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

I. INTRODUCTION

1.1 Background

The development of science and technology today is greatly influenced by the role of mathematics, which provides the basics of structure and logic in building sciences that are useful in everyday life. Over time, learning methods have progressed towards a more modern, effective, and efficient approach, one of which is by utilizing data mining techniques. Data mining is a series of processes to find patterns or important information from large-scale data sets through the application of certain methods. One of the methods in data mining is classification, which is the process of recognizing patterns or models that are able to describe and distinguish classes in a dataset. The main goal is for the model to predict the class label of unknown data. The model building process is based on the analysis of the training data. The model generated from the classification process can be used to classify new data or to predict future data trends (Widjiyati, 2021).

Classification can be used to solve various problems such as data classification, image classification, text categorization, structure prediction and others. An optimal classification model is able to describe the relationship pattern between data and its class label. The main problem in classification is how to create a model that is able to separate the data based on its class accurately (Fajri & Primajaya, 2023). The method commonly used for classification is Support Vector

Machine (SVM). Support Vector Machine (SVM) is a method that works based on the principle of finding the optimal hyperplane that serves as the boundary between classes by maximizing the distance or margin between these classes (Rantini et al., 2019).

One of the problems in building classification models is determining the most appropriate parameter values, as this has a significant impact on improving model performance. This process is known as hyperparameter tuning which makes it very important because hyperparameter values can have a significant impact on the model's ability to understand and generalize data properly (Alhakeem et al., 2022). Before performing the model training process, the hyperparameters must be determined first so as not to cause problems such as overfitting, underfitting or non-optimal model performance.

To overcome these problems, two common methods used to optimize hyperparameters are GridSearch and RandomSearch. GridSearch performs exhaustive search across the specified hyperparameter space, while RandomSearch randomly samples from that space. Each method has advantages and disadvantages, and choosing the right method can have a significant effect on the classification results.

Research related to hyperparameter optimization problems can be found in research conducted by Fajri & Primajaya (2023), where researchers successfully classify with the SVM method and get an average accuracy value of 84% from optimization results using GridSearch and RandomSearch. This shows that the accuracy value of GridSearch and RandomSearch has no significant difference. Another study was conducted by Putra & Pramartha (2025), where researchers successfully conducted classification to predictβ -Thalassemia disease using the SVM method. The results show that GridSearch hyperparameter optimization can determine the best parameters according to the value to be tested. In addition,

Fitra, et al. (2024), discussed the development of SVM models to improve the classification accuracy of heart disease diagnosis. The results showed that the accuracy increased from 87.65% to 96.56% from the optimization results using GridSearch. In the research of Rusman, et al (2023), discussing hyperparameter tuning optimization in the SVM method for classification of coffee fruit maturity levels. The research shows that classification using the SVM method gets an accuracy value of 99.44% from the optimization results using GridSearch. Then research conducted by Awalullaili, et al. (2022), discussed the classification of hypertension using the SVM GridSearch and SVM Genetic Algorithm (GA) methods. The results showed that classification with the SVM method got an accuracy of 89.42% with optimization using GridSearch.

The deadliest non-communicable disease in the world is heart disease. In Indonesia, heart disease is a serious threat to public health and occupies the top position in the list of causes of death, both in hospitals and outside health facilities. Some of the risk factors that affect the emergence of heart disease include hypertension, high cholesterol, diabetes, obesity and others. Symptoms of heart disease are often not realized by the sufferer until it reaches a fairly severe stage. Therefore, with the advancement of information technology, especially in the field of data science, there is an approach to help the diagnosis process. To predict whether someone is at risk of suffering from heart disease based on medical record data such as blood pressure, cholesterol, age, and other factors. By utilizing classification algorithms, heart disease detection systems can be developed to be more accurate, efficient and can help make medical decisions quickly.

Based on the description above, in this study the author is interested in optimizing hyperparameter tuning using GridSearch and RandomSearch to classify heart disease with the Support Vector Machine (SVM) method.

1.2 Research Objective

The objectives to be achieved in writing this report are :

1. Evaluate hyperparameter tuning optimization using GridSearch and RandomSearch in the SVM method in increase the classification accuracy of heart disease patients.
2. Comparing the performance of GridSearch and RandomSearch hyperparameter tunning optimization based on accuracy, precision, recall, and f1-score evaluation matrices.

1.3 Research Benefit

The benefits that can be taken from writing this report are :

1. Add insight for the author, especially regarding the application of the SVM method.
2. Knowing GridSearch and RandomSearch optimization to increase model performance in classifying heart disease patients.
3. Provide an overview to the public and government about heart disease.

II. LITERATURE REVIEW

2.1 Data Mining, Machine Learning, and Classification

Data Mining is a technique used to find patterns and trends in very large data sets using various techniques such as classification, association, clustering, prediction, and estimation (Han, et al., 2012). According to Arhami & Muhammad (2020), data mining is part of the Knowledge Discovery in Database (KDD) process, which serves to find patterns or information hidden in data. Data mining has two types of tasks: the first, descriptive tasks, which aim to explain the general characteristics of the data to make it easier to understand, and the second predictive tasks, which aim to create knowledge models that can be used to make predictions.

Machine Learning is a technique that improves system performance by learning from experience through algorithms that build models. This process aims to acquire new intelligence or knowledge by gradually increasing the ability to learn without explicit programming (Zhou, 2021). The machine learning process consists of two stages: the training stage and the testing stage. The training stage aims to train the algorithm by providing it with data, information, or experience in order for it to learn. to understand the patterns in the data. Meanwhile, the testing phase is used to evaluate the performance of the trained algorithm.

Classification is a process that consists of two main stages, namely the learning (training) stage and the classification stage. In the learning stage, a classification model is formed, while in the classification stage, the model is used to determine the class of the given data (Hamid & Hidayat, 2019). Figure 2.1 below shows the flow of stages in the classification process:

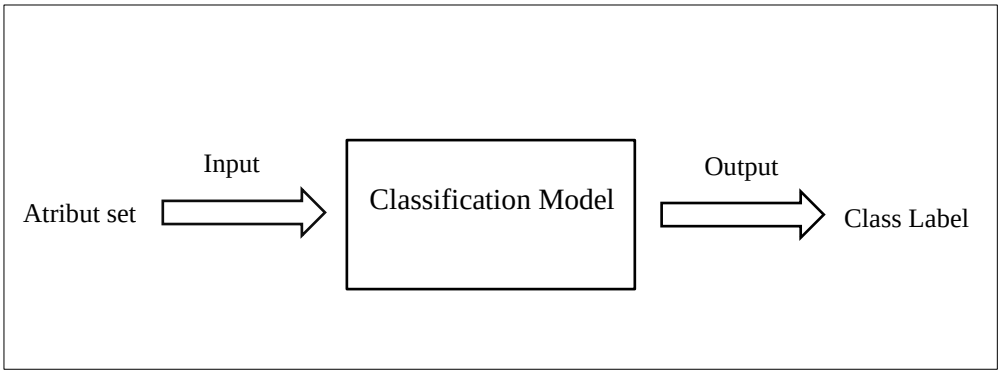

Figure 1. Classification Stages

In the training stage, the training data with known classes is analyzed using classification algorithms to build a model of each class. Since the training data has class labels, this process is referred to as supervised learning. The resulting model will be in the form of classification rules. These rules are then tested using test data, which is independent and not used during the training process, to assess their accuracy. If the test results show high accuracy, then the rules can be used to classify new data that has no known class.

2.2 Data Preprocessing

Data preprocessing is data that was originally a collection of raw and unstructured information processed systematically by selecting information that is considered important and relevant for analysis purposes. Furthermore, the data is transformed and rearranged so that it becomes a more organized and structured format, making it easier for further processing. This process is very important so that the data produced is not only cleaner and free from disturbances such as noise or irrelevant

data, but also so that the system can process the data more efficiently and accurately in producing the desired output (Aditya et al., 2022).

2.2.1 Balancing Data using Random Oversampling (ROS)

Data imbalance occurs when the distribution of classes in a dataset is unbalanced, that is, when the amount of data in one class (majority class) is much more than the other class which has less data (minority class) (Ali et al., 2013). This condition can cause the classification model to be less accurate because data from the minority class is often misclassified as the majority class. To overcome this imbalance problem, several methods can be applied, namely random oversampling.

Oversampling is a method that aims to increase the amount of data in the minority class to balance or approach the amount of data in the majority class (Chawla, 2009). One commonly used oversampling technique is Random Oversampling, which adds data by synthesizing or duplicating samples from the minority class randomly and repeatedly into the training data. Different from oversampling, undersampling removes data from the majority class until the amount of data in each class is equal. This process is carried out until the amount of data in the minority class is equal to the majority class (Aryanti et al., 2023).

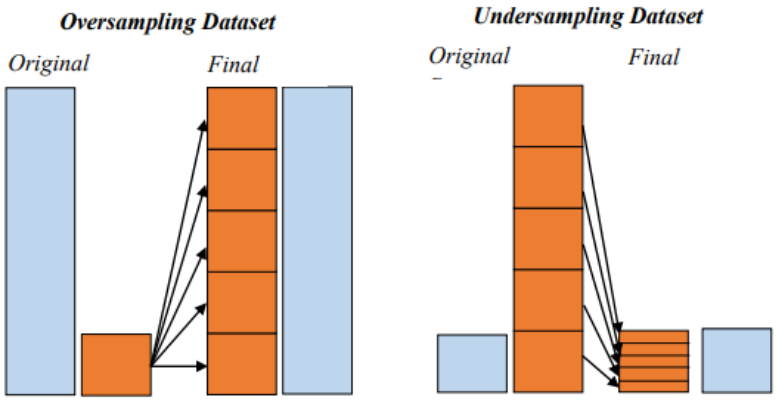

Figure 2. Oversampling and Undersampling Process

Figure 2. shows that the oversampling process is done by increasing the amount of data in the minority class using various methods, until the amount is equal to the majority class. On the other hand, undersampling is done by subtracting some data from the majority class to make it equal to the minority class.

2.2.2 Data Scaling

Data Scaling or normalization process is a method to equalize the numerical values in a dataset to a uniform scale without eliminating the differences in the range of values. This normalization plays a role in accelerating the learning process in machine learning algorithms (Li & Zhenyu, 2011). Therefore, the data from leaf venation feature extraction is normalized first.

a. Min-Max Normalization

   Min-max normalization resizes the data from the original range, so that all values are in the range of 0 and 1. The equation can be seen in equation (2.1).

$$v_{norm} = \left(\frac{v_i - v_{min}}{v_{max} - v_{min}}\right) \quad (2.1).$$

b. Standarization (Zero-Mean)

   The Zero-Mean normalization method is based on the mean and standard deviation. Standardizing a dataset involves rescaling the distribution of values, so that the observed mean is 0 and the standard deviation is 1. The standard deviation is calculated using equation (2.2).

$$x_{std} = \sqrt{\frac{1}{N-1}\sum_{I=1}^{N}(x_i - x_{mean})^2} \quad (2.2).$$

   x_{mean} is the average of the data. Normalization can be calculated by equation (2.3).

$$x_i' = \frac{x_i - x_{mean}}{x_{std}} \quad (2.3).$$

2.3 Support Vector Machine

Support Vector Machine (SVM) is a machine learning method that utilizes a high-dimensional feature space and uses linear functions as the basis for learning. The training process is based on algorithms rooted in optimization theory and integrates a learning bias. The SVM approach offers various advantages, such as model dependency on a small subset of data points known as support vectors, which also simplifies model interpretation. The main goal of SVM is to find the optimal hyperplane that can separate two classes of data in the input space. This hyperplane is a straight line in two-dimensional space and a flat plane in higher dimensions. SVM works by training the model using the training dataset, then generalizing to predict data that has never been seen before (Nurleli, 2023).

2.3.2 Concept SVM

SVM is a classification technique in machine learning that belongs to the supervised learning category, which works by predicting classes based on patterns obtained from the training process. SVM creates a hyperplane in an infinite high-dimensional space, which is used for both regression and classification purposes. This method utilizes linear functions in high-dimensional feature space based on statistical learning principles (Lestari & Sri, 2022).

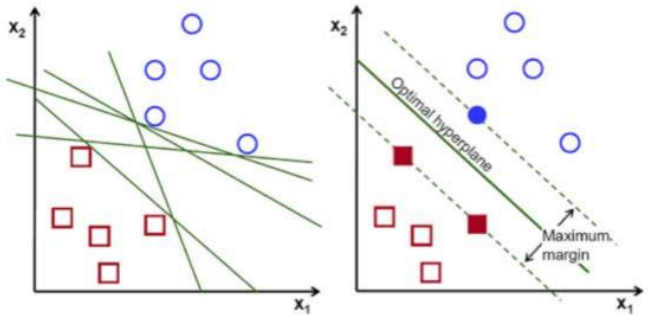

Figure 3. Maximum Margin in Hyperplane Determination

In Figure 3, we can see patterns that fall into two classes, namely+1 dan -1. Patterns belonging to the +1 class are marked with red squares, while patterns of the-1 class are marked with blue circles. The essence of the SVM method is to find the most optimal hyperplane to separate the two classes with the widest margin (Mohit, et al., 2021). The hyperplane itself is the dividing line between two groups of data, while the margin is the distance between the hyperplane and the closest data point from each class. Two bounding planes are used to mark the boundaries of the first class and the second class separately. The data points closest to this best hyperplane are called support vectors.

For classification cases with more than two classes, special strategies such as One-Against-One (OAO) and One-Against-All (OAA) methods are required. The OAO method works by building as many binary SVM models as class pair combinations, i.e.$k!/(2!\,(k-2))$, where each model is trained using data from two different classes. While the OAA method creates k binary SVM models (k is the number of classes), where each model to $-i$ is trained using all data to find a solution to the problem (Nurkholis, et al., 2022).
Suppose the available data is represented as a vector:

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}; x_i \in R, y_i \in \{-1,1\} \quad (2.4)$$

A dataset is given the variable x_i, while the classes in the dataset are represented by the variable y_i. The first class separated by the hyperplane is assigned a value of 1 and the other classes are assigned a value of -1.

So that the following equation is obtained:

$$\boldsymbol{w}.\boldsymbol{x_i} + b = 0 \tag{2.5}$$

with:

$\boldsymbol{w}$ = the weight value of the support vector perpendicular to the hyperplane

b = bias value

Then the following equation is obtained:

$$\boldsymbol{w}.\boldsymbol{x_i} + b \geq +1 \text{ for } y_i = +1 \tag{2.6}$$

$$\boldsymbol{w}.\boldsymbol{x_i} + b \leq -1 \text{ for } y_i = -1 \tag{2.7}$$

with:

$\boldsymbol{x_i}$ = vector feature data to i (data input),

y_i = class data label to i.

To achieve the maximum margin between classes, we maximize the distance between the hyperplane and the data pattern. The margin is defined as $d = d_1 + d_2$, so the margin will reach the maximum value if $d_1 = d_2$. Finding the largest margin value is done by maximizing the distance between the hyperplane and its closest, which can be measured as to $\frac{1}{\|\boldsymbol{w}\|}$.

$$d = d_1 + d_2 = \frac{1}{\|\boldsymbol{w}\|} + \frac{1}{\|\boldsymbol{w}\|} = \frac{2}{\|\boldsymbol{w}\|} \tag{2.8}$$

Referring to the equation above, to get the maximum margin is equal to minimizing the value of $\|\boldsymbol{w}\|^2$, systematically stated as follows.

$$\min \tau(w) = \frac{1}{2}\|\boldsymbol{w}\|^2 \tag{2.9}$$

Then, optimization can be done by applying the Lagrange multiplier as follows.

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{l} a_i[y_i(\boldsymbol{w}.\boldsymbol{x_i} + b) - 1]$$

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{l} a_i y_i(\boldsymbol{w}.\boldsymbol{x_i} + b) - \sum_{i=1}^{l} a_i \tag{2.10}$$

a_i is a Lagrange multiplier with zero or positive values ($a_i \geq 0$). The optimization process is done by minimizing L against w and b, as follows.

$$\frac{\partial L}{\partial b} = 0$$

$$\sum_{i=1}^{l} a_i y_i = 0 \tag{2.11}$$

$$\frac{\partial L}{\partial w} = 0$$

$$\boldsymbol{w} - \sum_{i=1}^{l} a_i y_i \boldsymbol{x_i} = 0$$

$$\boldsymbol{w} = \sum_{i=1}^{l} a_i y_i \boldsymbol{x_i} \tag{2.12}$$

In addition, to perform optimization, it can be done by maximizing L againts a_i by substituting equations (2.8) and (2.9) into equation (2.6) as follows.

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{l} a_i y_i (\boldsymbol{w}.\boldsymbol{x_i} + b) - \sum_{i=1}^{l} a_i$$

$$L = \frac{1}{2}(\boldsymbol{w}.\boldsymbol{w}) - \left(\sum_{i=1}^{l} a_i y_i \boldsymbol{w}.\boldsymbol{x_i} + \sum_{i=1}^{l} a_i y_i b - \sum_{i=1}^{l} a_i\right)$$

$$L = \frac{1}{2}\left(\sum_{i=1}^{l} a_i y_i \boldsymbol{x_i} . \sum_{i=1}^{l} a_j y_j \boldsymbol{x_j}\right) - \left(\left(\sum_{i=1}^{l} a_i y_i \boldsymbol{x_i} . \sum_{i=1}^{l} a_j y_j \boldsymbol{x_j}\right) + 0 - \sum_{i=1}^{l} a_i\right)$$

$$L = \frac{1}{2}\sum_{i=1}^{l}\sum_{i=1}^{l} a_i a_j y_i y_j \boldsymbol{x_i x_j} - \left(\sum_{i=1}^{l}\sum_{i=1}^{l} a_i a_j y_i y_j \boldsymbol{x_i x_j} - \sum_{i=1}^{l} a_i\right)$$

$$L = \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{i=1}^{l} a_i a_j y_i y_j \boldsymbol{x_i x_j} \tag{2.13}$$

Where, $a_i \geq 0, \sum_{i=1}^{l} a_i\, y_i = 0$

The value of a_i can be calculated by solving equation (2.10), which is used to find the primal variable using the formula:

$$\boldsymbol{w} = \sum_{i=1}^{l} a_i\, y_i K(\boldsymbol{x_i x_j}), b = -\frac{1}{2}(\boldsymbol{w}.\boldsymbol{x}^+ + \boldsymbol{w}.\boldsymbol{x}^-) \tag{2.14}$$

Then the value a_i is obtained which is called the support vector, while the others have the value $a_i = 0$. The resulting decision function is only influenced by the value of the support vector.

2.3.2 Kernel Function

The SVM algorithm utilizes a number of mathematical functions called kernels to process data. These kernels accept data as input and then transform it into another form according to the needs of the model. The selection of the right kernel

function is an important factor in the performance of classification using SVM. Some types of kernel functions that are often used include linear, polynomial (poly), and Radial Basis Function (RBF).

a. Linear Kernel

Linear kernel is the simplest type of kernel function. It is used when the data being analyzed is already linearly separable. Linear kernels are suitable for data with many features because mapping to a higher dimensional space does not always improve performance, as is the case in text classification. In text classification, both the amount of data (documents) and the number of features (words) are usually very large (Kowalczyk, 2014). The following is an example of using a linear kernel in SVM.

$$K(x, y) = x.y \qquad (2.15).$$

the mapping function ϕ is the identity / there is no mapping.

b. Polynomial Kernel

A polynomial kernel is a type of kernel function that is used when the data is not linearly separable. This kernel is particularly suitable for cases where the entire training data has gone through a normalization process.

$$K(x, y) = (x.y + 1)^p \qquad (2.16).$$

c. Radial Basis Function (RBF) Kernel

RBF kernel is one of the commonly used kernel functions in the analysis of linearly non-separable data. This kernel has two main parameters, namely Gamma and Cost (C)**.** The C parameter plays a role in regulating how much the model prioritizes reducing misclassification in the training data, thus helping to optimize SVM performance. Meanwhile, Gamma determines the extent to which a data point influences the formation of the dividing line. A low gamma

value indicates that points far from the dividing line still have an influence, while a high gamma value indicates that only points that are very close to the dividing line are taken into account. The following is the general form of the RBF kernel equation.

$$K(x, y) = \exp\left[-\gamma \|x, y\|^2\right. \quad (2.17).$$

2.4 GridSearch dan RandomSearch

GridSearch is one of the methods used to determine the best parameters in a model. This method works by testing every possible combination of parameters based on values that have been specified by the user. All combinations are arranged in a grid, and the optimal parameters are selected based on the combination that produces the smallest error (Saputra, et al. 2019). In contrast to GridSearch which tests all available combinations, RandomSearch selects a number of combinations randomly from the parameter space. The main focus of RandomSearch is to explore the parameters that have the most influence on model performance (Valarmathi & Sheela, 2021). Determining the best model resulting from hyperparameter tuning using GridSearch and RandomSearch is done by comparing model performance evaluation values, such as accuracy, precision, recall, and f1-score. The model that produces the highest and most consistent evaluation value is considered the optimal model in performing classification.

2.5 Evaluation Model

After building both the basic and enhanced versions of the SVM model, the next step is to test the performance of the model. This test produces an evaluation in the form of a classification report, which is used to measure the predictive quality of the classification algorithm. Classification report consists of four main metrics, namely accuracy, precision, recall, and F1-score. Accuracy indicates the frequency of the model's success in correctly classifying heart disease. Precision

measures the level of accuracy of the model in generating correct positive predictions, by minimizing prediction errors. Recall describes the extent to which the model is able to recognize all true positive cases, which is the ratio of correct positive predictions compared to the total positive data. The F1-score is a harmonic mean between precision and recall, reflecting the balance between the two. Meanwhile, confusion matrix is used as an evaluation method to assess the performance of machine learning algorithms (Terrada, et al. 2019).

Table 1. Confusion Matrix Table

	Predictive Positive	**Predictive Negative**
Actual Positive	TP	FP
Actual Negative	FN	TN

Table 1. shows the confusion matrix for 2-class classification. True Negative (TN) indicates the number of data points in the negative class that were correctly classified, False Positive (FP) indicates the number of data points that are actually negative but were incorrectly classified as positive, False Negative (FN) indicates the number of data points that are actually positive but were incorrectly classified as negative, and True Positive (TP) indicates the number of data points in the positive class that were correctly classified (Wardhani et al., 2019).

Evaluation with confusion matrix produces accuracy, precision, recall, and f1- score values as follows:

1. Accuracy is the comparison value between the correctly classified data and the entire data. High accuracy indicates that the model has a good ability to classify data. Accuracy is systematically expressed in the following equation (2.11).

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (2.18)$$

2. Precision is the ratio of positive data classified as correct to the total number of positive prediction results. Precision is mathematically expressed in the following equation (2.12).

$$Precision = \frac{(TP)}{(TP+FP)} \quad (2.19)$$

3. Recall is the ratio of the number of correctly predicted positive data (TP) to the actual number of positive data. Recall is mathematically expressed in equation (2.13).

$$Recall = \frac{(TP)}{(TP+FN)} \quad (2.20)$$

4. F1-score is a balance ratio between precision and recall. F1-score is mathematically expressed in the following equation (2.13).

$$F1 - score = \frac{2(Precision \times Recall)}{(Precision \times Recall)} \quad (2.21)$$

III. RESEARCH METODOLOGY

3.1 Time and Place of Research

This research was conducted in the odd semester of the 2024/2025 academic year at the Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Lampung.

3.2 Research Data

The data used in this study are secondary data, namely heart disease data obtained from the official kaggle website which can be accessed (https://www.kagle.com). The data consists of 918 data with independent variables, namely Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, and ST_Slope, and one dependent variable, namely category / HeartDisease.

The explanation of each variable is shown in Table 2 below

Table 2. Research Variables

Variables	Variable Definition	Data Type
HeartDisease (Y)	Target/label, whether the patient has heart disease	Categorical (Binary): (0= No) (1= Yes)
Age (X_1)	Patient age in years	Numeric (Integer)
Sex (X_2)	Patient gender	Categorical: (0= Male) (1= Female)
ChestPainType(X_3)	Types of chest pain	Categorical: (1=ATA) (2=NAP) (3=ASY) (4=TA)
RestingBP (X_4)	Blood pressure at rest	Numeric (Integer)
Cholesterol (X_5)	Blood cholesterol level (mg/dL)	Numeric (Integer)
FastingBS (X_6)	Fasting blood sugar >120 mg/dl	Categorical (Binary): (0= No) (1= Yes)
RestingECG (X_7)	Electrocardiogram results at rest	Categorical : (1=Normal) (2=ST) (3=LVH)
MaxHR (X_8)	Maximum heart rate achieved during exercise	Numeric (Integer)
ExerciseAngina (X_9)	Does the patient experience angina during exercise	Categorical (Binary): (0= No) (1= Yes)
Oldpeak (X_{10})	Exercise-induced ST depression compared to Rest	Numeric (Float)
ST_Slope(X_{11})	ST segment slope	Categorical (1=UP) (2=Flat) (3=Down)

3.3 Research Method

The steps taken in this research are as follows:

1. Perform data visualization with pie charts to show the percentage of people with heart disease and perform descriptive analysis to see a summary of the data characteristics of the research variables by evaluating the average value.
2. Perform data preprocessing:
   a. Cleaning the data to see if there are missing values or duplicated rows.
   b. Performing variable selection to remove irrelevant features.
3. Handle imbalance data with ROS.
4. Scaling the data to transform the data using a standard scaler.
5. Divide data into training data and testing data with split data 90% training data and 10% testing data.
6. Perform classification process using SVM, by modeling both with default settings and with hyperparameter tuning optimization using GridSearch and RandomSearch.
7. Evaluate SVM classification results and compare the performance of GridSearch and RandomSearch hyperparameter tunning optimization based on accuracy, precision, recall, and f1-score evaluation matrices in classifying heart disease.

V. CONCLUSION

Based on the results of the research that has been done, the following conclusions can be obtained:

1. Optimizing hyperparameter tuning using RandomSearch, as well as applying balancing techniques with ROS, has succeeded in improving the performance of the SVM model in classifying heart disease. This can be seen from the increase in accuracy value from 85.29% to 89.21%.
2. Hyperparameter tuning using GridSearch, as well as the application of balancing techniques with ROS, does not provide changes to the performance of the SVM model, indicated by a fixed accuracy value of 85.29%.
3. Based on the comparison of GridSearch and RandomSearch hyperparameter tuning optimization with ROS performance based on accuracy, precision, recall, and f1-score evaluation matrices, RandomSearch hyperparameter tuning optimization shows better results in classifying heart disease.

BIBLIOGRAPHY

Aditiya, P., Enri, U., & Maulana, I. 2022. Sentiment Analysis of Myim3 User Reviews on the Google Play Site Using Support Vector Machine. *Journal Ris. Computer (*JURIKOM). **9**(4):1020.

Alhakeem, Z.M., Yasir, M.J., Sadiq, N.H., Hamzah, I., Luis, F.A.B., & Hussein, M.H. 2022. Predicting of Ecofriendly Concert Compressive Strength using Gradient Boosting Regression Tree Combined with GridSearchCV Hyperparameter Optimization Technique. *Materials*. **15**(21):7432.

Ali, A., Siti, M.S.,& Anca, L.R.2015. Classification with Class Imbalance Problem: A Review. *International Journal Advance Soft Computer Application*. **7**(3): 176-204.

Arhami, M. & Muhammad, N.2020. *Data Mining: Algorithms and Implementation*. Andi Publisher, Aceh.

Aryanti, R., Titik, M., & Rahmat, H. 2023. Maternal Health Risk Classification using Random Oversampling to Overcome Data Imbalance. *KLIK*. **3**(5): 409-416.

Awalullaili, F., Ispriyanti, D., & Widiharih, T. 2022. Classification of Hypertension Disease Using SVM GridSearch and SVM Genetic Alghoritm (GA) Methods. *Journal Gaussian*. **11**(4):488-498.

Chawla, N.V.2009. *Data Mining for Imbalance Datasets: an Overview Data Mining and Knowledge Discovery Handbook*. Springer, Berlian.

Fajri, M., & Primajaya, A. 2023. Comparison of Hyperparameter Optimization Techniques in SVM for Classification Problems Using GridSearch and RandomSearch. *Journal of Applied Informatics and Computing* (JAIC). **7**(1): 10-15.

Fitra, G., Suroso., & Soim, S. 2024. Development of Support Vectoe Machine Model to Improve Classification Accuracy of Heart Disease Diagnosis. *Journal Technology System Information and Apliccation*. **7**(3): 1418-1428.

H. D. A. Hamid., & Hidayat, N. 2019 Diagnosis of Chili Plant Diseases Using the Modified K- Nearest Neighbor (MKNN) Method. *Journal of Infrastructure Technology Development and Computer Science*. **3**(3): 2881-2886.

Han, J., Kamber, M.,& Pei, J. 2012. *Data Mining: Concept and Techniques*. 3^{rd}Edition. Elsevier, San Francisco.

Kowalczyk, A. 2017. *Support Vector Machines Succinctly*. USA: Syncfusion.

Lestari, W. & Sri, S. 2022. Implementation of K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) for Classification Cardiovascular Disease. *Multiscience*. **2**(10): 30-36.

Li, W. & Zhenyu, L. 2011. A Method Of SVM With Normalization in Intrusion Detection. *Procedia Environmental Science*. **11**: 256– 262.

Mohit, I., Santhosh, K., Avula, U.K.R., & Badhagouni, S.K. 2021. An Approach to Detect Multiple Disease using Machine Learning Algorithm. *Jurnal of Physics: Conference Series*. **2089**(1): 1-7.

Mustaqim, M., Warsito, B & Surarso, B. 2019. Combination of Synthetic Minority Oversampling Technique (SMOTE) and Neural Network Backpropagation to Handle Unbalanced Data on Prediction of Implant Contraceptive Usage. *Scientific Journal of Information Systems Technology*. **5** (34) : 116–127.

Nurkholis, A., Debby, A.,&Aris, M. 2022. Comparison of Kernel Support Vector Machine Multi-Class in PPKM Sentiment Analysis on Twitter. *Jurnal Rekayasa Sistem dan Teknologi Informasi*. **6**(2):227-233.

Nurleli. 2023. *Application of Support Vector Machine (SVM) Method for Single Tuition Classification at North Sumatra State Islamic University*. Thesis. State Islamic University of North Sumatra

Putra, I., & Pramartha, C. 2025. Hyperparameter Optimization of Support Vector Machine Algorithm in β-Thalassemia Disease Classification. *National Journal of Information Technology and Its Applications*. **3**(2): 283-294.

Qu, Z., Li, H., Wang, Y., Zhang, J., Abu-Siada, A., & Yao, Y., 2020, Detection of Electricity Theft Behavior Based on Technique and Random Forest Classifier. *Energies*. **13** (8): 2039.

Rantini, D., Rosyida, I., &Santi, W.P. 2019. Predicting Popularity of Movie using Support Vector Machines. *INFERENSI.* **2(**1): 13-17.

Rusman, J., Haryati, B., & Michael, A.2023. Hyperparameter Optimization in Support Vector Machine Method for Classification of Coffee Fruit Maturity Level. *Journal of Informatics and Computers*. **11**(2): 195-202.

Saputra, G., Wigena, A., & Sartono, B. 2019. Use of Support Vector Regression in Modeling the Indonesian Sharia Stock Index with Grid Search Algorithm. *Journal. Statistics ITS*. **3**(2): 148-160.

Terrada., Oumaima., Cherradi, B., Raihani, A., & Bouattane, O. 2019. Classification and Prediction of Atherosclerosis Diseases Using Machine Learning Algorithms. *International Conference on Optimization and Applications (ICOA).* 1-5.

Valarmathi, R., & Sheela, T. 2021. Heart disease prediction using hyper parameter optimization (HPO) tuning. *Biomed Signal Process Control*. **70**.

Widjiyati, N. 2021. Implementation of the Random Forest Algorithm on Clasification Dataset Credit Approval. *Jurnal Janitra Informatika and Sistem Information*. **1**(1):1-7.

Zhou, Z.H. 2021. *Machine Learning*. Springer Nature, China.