



UNDERGRADUATE THESIS

Implementation of PSO-Random Forest in Classification of Heart Disease with Boruta Feature Selection and Data Balancing Technique SMOTE-Tomek

Faisal Wildan Habibi
NPM 21081010216

THESIS ADVISORS

Henni Endah Wahanani, S.T., M.Kom.
Achmad Junaidi, S.Kom., M.Kom.

**MINISTRY OF HIGHER EDUCATION, SCIENCE, AND TECHNOLOGY
UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR
FACULTY OF COMPUTER SCIENCE
INFORMATICS STUDY PROGRAM
SURABAYA
2026**

APPROVAL SHEET

**Application of PSO-Random Forest in Heart Disease Classification with
Boruta Feature Selection and SMOTE-Tomek Data Balancing Technique**

By :
FAISAL WILDAN HABIBI
NPM. 21081010216

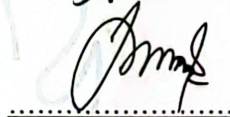
Has been defended before and accepted by the Thesis Examining Team of the Faculty of
Computer Science Study Program, Veteran National Development University, East Java on
February 11, 2026

Henni Endah Wahanani, ST, M.Kom.
NIP. 19780922 2021212 005



(Advisor I)

Achmad Junaidi, S.Kom., M.Kom.
NPT 3 7811 04 0199 1



(Advisor II)

Eka Prakarsa Mandyartha, ST, M.Kom.
NIP. 19880525 2018031 001



(Head Assessor)

Chrystia Aji Putra, S.Kom., MT
NIP. 19861008 2021211 001



(Assessor I)

Acknowledge by,
Dean of the Faculty of Computer Science


Prof. Dr. Ir. Novirina Hendrasarie, MT
NIP. 19681126 199403 2 001

APPROVAL SHEET

**Application of PSO-Random Forest in Heart Disease Classification with
Boruta Feature Selection and SMOTE-Tomek Data Balancing Technique**

By:
FAISAL WILDAN HABIBI
NPM. 21081010216

Approved to proceed to the Thesis Examination



Approved by,
Coordinator of Informatics Study Program
Faculty of Computer Science

Dr. Intan Yuniar Purbasari, S.Kom., M.Sc.
NIP. 19800602 2025212 029

STATEMENT OF ORIGINALITY

The undersigned below:

Student Name : FAISAL WILDAN HABIBI
Study Program : Informatics
Advisors : Henni Endah Wahanani, ST, M.Kom.
Achmad Junaidi, S.Kom., M.Kom.

hereby declare that the contents of part or all of the dissertation with the title:

**Application of PSO-Random Forest in Heart Disease Classification
with Boruta Feature Selection and SMOTE-Tomek Data Balancing
Technique**

is truly my own intellectual work, completed without use of any unauthorized materials, and not the work of another party, which I acknowledge as my own. All references that cited or referred are fully documented in the bibliography. If this statement proves to be untrue, I am willing to accept any sanctions in accordance with applicable regulations.

Surabaya, July 10 2026
Declarant,



Faisal Wildan Habibi
NPM. 21081010216

ABSTRACT

Student Name / NPM	Faisal Wildan Habibi / 21081010216
Thesis title :	Application of PSO-Random Forest in Heart Disease Classification with Boruta Feature Selection and SMOTE-Tomek Data Balancing Technique
Advisor :	1. Henni Endah Wahanani, ST, M.Kom. 2. Achmad Junaidi, S.Kom, M.Kom

Heart disease is the major cause of death globally with 17 million annual cases, demanding accurate early detection. This study proposes a PSO-Random Forest approach with the application of Boruta feature selection and SMOTE-Tomek balancing on the UCI Heart Disease dataset (920 samples, 16 features, multi-class 0-4). The methodology includes preprocessing consisting of median/mode imputation, IQR capping, and encoding, PSO optimization with `n_estimators` parameters of 400-500 and `max_depth` of 10-20, and 4 scenarios: baseline, SMOTE-Tomek, Boruta, hybrid. The evaluation results show that the application of Boruta+SMOTE-Tomek achieves an optimal F1-score of 0.5917 (accuracy 0.6087, precision 0.6145), outperforming the baseline of 0.5894. Boruta selected 8 key features (age, `trestbps`, `chol`, `thalach`, `dominant oldpeak`), SMOTE-Tomek balanced the minority class (0:55%→50%). The 80:20 binary classification ratio scenario achieved 83.15% accuracy, and the multi-class scenario achieved 57.61%. PSO-RF proved robust overfitting compared to standard RF, with Boruta reducing noise by 48% and SMOTE-Tomek increasing minority recall by 12%. This approach contributes practically to ML-based cardiac diagnostic systems supporting early detection of disease stages in medicine.

Keywords: PSO-Random Forest, Boruta, SMOTE-Tomek, Multi-Class Classification, UCI Heart Disease

FOREWORD

All praises only be to Allah SWT for all His grace, guidance and permissions to the author so that the thesis entitled "**Application of PSO-Random Forest in Heart Disease Classification with Boruta Feature Selection and SMOTE-Tomek Data Balancing Technique**" can be resolved well.

The author also would like to thank Mrs. Henni Endah Wahanani, ST, M.Kom. as the main advisor and Mr. Achmad Junaidi, S.Kom., M.Kom. as secondary advisor, who have taken their time to provide guidance, knowledge, advice, and motivations to the author. The author has also received more assistance from various parties, such as moral, spiritual, and material supports. Therefore, the author also would like to express his gratitude to:

1. Mrs. Prof. Dr. Ir. Novirina Hendrasarie, M.T. as Dean of the Faculty of Computer Science, National Development University "Veteran" East Java.
2. Mrs. Dr. Intan Yuniar Purbasari, S.Kom., M.Sc. as coordinator study program Informatics Faculty of Social Sciences and Computer Science, National Development University "Veteran" East Java.
3. Mrs. Henni Endah Wahanani, ST, M.Kom. as the main advisor.
4. Mr. Achmad Junaidi, S.Kom, M.Kom. as the secondary advisor.
5. All Informatics Study Program Lecturers as a whole.

The author realizes that this thesis may contain many shortcomings. Therefore, any of constructive criticism and suggestions from all parties are will be highly appreciated for the perfection of this thesis. Finally, despite any of its limitations, the author truly hopes that this report will be beneficial for all parties in general and the author in particular.

Surabaya, July 10 2026

Author,

Faisal Wildan Habibi

21081010216

LIST OF CONTENTS

APPROVAL SHEET	3
APPROVAL SHEET	5
STATEMENT OF ORIGINALITY	7
ABSTRACT	9
FOREWORD.....	11
LIST OF CONTENTS	13
LIST OF FIGURES	15
LIST OF TABLES	17
CHAPTER I INTRODUCTION.....	1
1.1 Background	1
1.2 Problem Formulation.....	3
1.3 Research Objectives	3
1.4 Problem Definition.....	4
1.5 Benefits of Research.....	4
CHAPTER II LITERATURE REVIEW	5
2.1 Previous Research	5
2.2 Heart Disease.....	9
2.2.1 Risk Factors for Heart Disease.....	9
2.2.2 Difficulty in Detecting Heart Disease	11
2.3 Random Forest	11
2.4 Particle Swarm Optimization	13
2.5 Boruta Feature Selection	15
2.6 SMOTE-Tomek.....	17
2.7 Evaluation Metrics	18
CHAPTER III SYSTEM DESIGN AND IMPLEMENTATION	21
3.1 Research Methodology.....	21
3.2 Research Stages.....	21
3.2.1 Literature Study.....	21
3.2.2 Data Collection.....	22
3.2.3 Data Pre-Processing	23
3.2.4 Model Creation.....	28
3.2.4.1 Baseline (RF-PSO).....	28
3.2.4.2 Implementation of SMOTE-Tomek Data Balancing	29
3.2.4.3 Implementation of Boruta Feature Selection	30
3.2.5 Model Evaluation	32
3.2.6 Testing Scenario.....	32
3.2.7 Analysis of Results.....	33
CHAPTER IV TESTING AND ANALYSIS	35

4.1 Data Pre-Processing	35
4.1.1 Data Exploration and Analysis (EDA).....	35
4.1.2 Data Cleaning	39
4.1.3 Data Transformation	43
4.1.4 Data Sharing	46
4.2 Model Creation.....	46
4.2.1 Baseline	47
4.2.2 Implementation of Smote-Tomek Data Balancing.....	47
4.2.3 Implementation of Boruta Feature Selection	48
4.2.4 Application of All Techniques	49
4.3 Model Evaluation	49
4.4 Testing Scenario	55
4.5 Analysis of Results.....	58
CHAPTER V CLOSING	61
5.1 Conclusion.....	61
5.2 Suggestions.....	61
BIBLIOGRAPHY	63
APPENDIX	65

LIST OF FIGURES

CHAPTER II LITERATURE REVIEW	6
2.3 Random Forest	12
Figure 2.3.1 Random Forest Algorithm Flowchart	13
2.4 Particle Swarm Optimization	14
Figure 2.4.1 PSO Algorithm Flowchart	15
2.5 Boruta Feature Selection	16
Figure 2.5.1. Boruta Algorithm Flowchart.....	17
2.6 SMOTE-Tomek.....	18
Figure 2.6.1 Flowchart of the SMOTE-Tomek Algorithm	18
CHAPTER III SYSTEM DESIGN AND IMPLEMENTATION.....	21
3.2 Research Stages.....	21
Figure 3.2.1 Research Stages Diagram	21
3.2.3 Data Pre-Processing	23
Figure 3.2.3.1 Data Preprocessing Flowchart	23
Figure 3.2.3.2 EDA Results of Heart Disease Data	24
3.2.4 Model Creation.....	28
Figure 3.2.4.1 Model Creation Flowchart.....	28
3.2.4.1 Baseline (RF-PSO).....	28
Figure 3.2.4.1 RF-PSO Flowchart.....	29
3.2.4.2 Implementation of SMOTE-Tomek Data Balancing	29
Figure 3.2.4.2.1 SMOTE-Tomek Flowchart	29
3.2.4.3 Implementation of Boruta Feature Selection	30
Figure 3.2.4.3.1 Boruta Flowchart	31
CHAPTER IV TESTING AND ANALYSIS	34
4.1 Data Pre-Processing	34
4.1.1 Data Exploration and Analysis (EDA).....	34
Figure 4.1.1.1 Data Structure and Content.....	34
Figure 4.1.1.2 Empty Values in Data	35
Figure 4.1.1.3 Numeric Column Statistics	35
Figure 4.1.1.4 Distribution of Numerical Data	36
Figure 4.1.1.5 Distribution of Categorical Data.....	37
4.1.2 Data Cleaning.....	38
Figure 4.1.2.1 Data Before Column Deletion	39
Figure 4.1.2.2 Data After Column Deletion.....	39
Figure 4.1.2.3 Data Before Handling Empty Values	40
Figure 4.1.2.4 Data After Handling Empty Values.....	41
Figure 4.1.2.5 Data Before Outlier Handling.....	42
Figure 4.1.2.6 Data After Outlier Handling	42

4.1.3 Data Transformation	42
Figure 4.1.3.1 Data Before Transformation	44
Figure 4.1.3.2 Data After Transformation.....	44
4.1.4 Data Sharing.....	45
Figure 4.1.4.1 Data Division Results	45
4.2 Model Creation.....	45
4.2.1 Baseline	46
Figure 4.2.1.1 Baseline Training Results	46
4.2.2 Implementation of Smote-Tomek Data Balancing.....	46
Figure 4.2.2.1 Results of Implementing Smote-Tomek	47
4.2.3 Implementation of Boruta Feature Selection	47
Figure 4.2.3.1 Results of Boruta Implementation	48
4.2.4 Application of All Techniques	48
Figure 4.2.4.1 Results of Applying All Techniques.....	48
4.3 Model Evaluation	49
Figure 4.3.1 Confusion Matrix Baseline	50
Figure 4.3.2 Baseline Metrics	50
Figure 4.3.3 Confusion Matrix Application of Smote-Tomek.....	51
Figure 4.3.4 Smote-Tomek Implementation Metrics	51
Figure 4.3.5 Confusion Matrix of Boruta Application.....	52
Figure 4.3.6 Evaluation Metrics of Boruta Implementation	53
Figure 4.3.7 Confusion Matrix Application of All Techniques	53
Figure 4.3.8 Evaluation Metrics for Implementation of All Techniques	54

LIST OF TABLES

CHAPTER II LITERATURE REVIEW	6
2.7 Evaluation Metrics	19
Table 2.7.1 Confusion Matrix Table	19
CHAPTER III SYSTEM DESIGN AND IMPLEMENTATION	21
3.2.2 Data Collection	22
Table 3.2.2.1 Dataset Attribute Details	22
Table 3.2.2.2 Example of Heart Disease Dataset Sample	23
3.2.3 Data Pre-Processing	23
Table 3.2.3.1 Example of Checking Empty Values	24
Table 3.2.3.2 Example of Numerical Data Distribution	24
Table 3.2.3.3 Example of Categorical Data Distribution	25
Table 3.2.3.4 Data Table Before Cleaning	26
Table 3.2.3.5 Data Table After Cleaning	26
Table 3.2.3.6 Data Table Before Categorical Handling	26
Table 3.2.3.7 Data Table After Categorical Handling	27
3.2.4 Model Creation	28
3.2.4.2 Implementation of SMOTE-Tomek Data Balancing	29
Table 3.2.4.4.1 Distribution Table Before Balancing	30
Table 3.2.4.2.2 Distribution Table After Balancing	30
3.2.4.3 Implementation of Boruta Feature Selection	30
Table 3.2.4.3.1 Feature Table Before Selection	31
Table 3.2.4.3.2 Feature Table After Selection	31
Table 3.2.4.1 Model Creation Parameter Table	31
3.2.6 Testing Scenario	32
Table 3.2.6.1 Test Scenario Parameter Table	32
3.2.7 Analysis of Results	33
Table 3.2.7.1 Evaluation Table of Analysis Results	33
CHAPTER IV TESTING AND ANALYSIS	34
4.1 Data Pre-Processing	34
4.1.2 Data Cleaning	38
Table 4.1.2.1 Number and Percentage of Blank Values	38
4.1.3 Data Transformation	42
Table 4.1.3.1 Column Type Details	43
4.4 Testing Scenario	54
Table 4.4.1 Data splitting ratio 70:30	54
Table 4.4.2 Data splitting ratio 75:25	55
Table 4.4.3 Data splitting ratio 80:20	55
Table 4.4.4 Binary Ratio 70:30	55
Table 4.4.5 Binary Ratio 75:25	55

Table 4.4.6 Binary Ratio 80:20.....	56
4.5 Analysis of Results.....	57
Table 4.5.1 Optimal Model per Ratio	57