



UNDERGRADUATE THESIS

APPLICATION OF THE CATBOOST ALGORITHM WITH OPTUNA OPTIMIZATION FOR CLASSIFYING THE ETIOLOGY OF ACUTE FEVER USING SMOTE-NC AS A DATA CLASS BALANCER

RAIHAN RAMADHAN
NPM 22081010151

THESIS ADVISORS

Yisti Vita Via, S.ST., M.Kom.

Ardhon Rakhmadi, S.Tr.T., 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 THE CATBOOST ALGORITHM WITH OPTUNA OPTIMIZATION FOR CLASSIFYING THE ETIOLOGY OF ACUTE FEVER USING SMOTE-NC AS A DATA CLASS BALANCER

By:
RAIHAN RAMADHAN
NPM. 22081010151

Has been defended before, and accepted by, the Board of Assessors of the Thesis Examination of the Informatics Study Program, Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jawa Timur, on August 20, 2026:

Approved,

Yisti Vita Via, S.ST., M.Kom
NIP. 19860425 202121 2 001



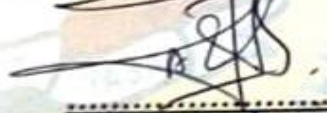
(Advisor I)

Ardhon Rakhmadi, S.Tr.T., M.Kom.
NIP. 19910805 202406 1 002



(Advisor II)

Fetty Tri Anggraeny, S.Kom, M.Kom
NIP. 19820211 202121 2 005



(Head Assessor)

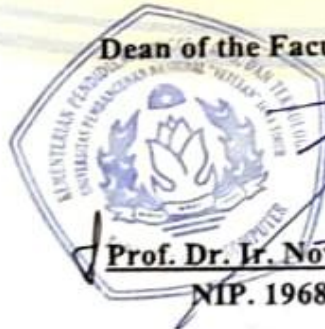
Eva Yulia Puspaningrum, S.Kom., M.Kom
NIP. 19890705 202121 2 002



(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 THE CATBOOST ALGORITHM WITH OPTUNA
OPTIMIZATION FOR CLASSIFYING THE ETIOLOGY OF ACUTE
FEVER USING SMOTE-NC AS A DATA CLASS BALANCER**

By:
RAIHAN RAMADHAN
NPM. 22081010151

Approved to proceed to the Thesis Examination

Approved by,

Coordinator of Informatics Study Program
Faculty of Computer Science



Dr. Intan Yuniar Purbasari, S.Kom. MSc.

NIP. 19800602 202521 2 029

STATEMENT OF ORIGINALITY

I am the undersigned:

Student Name : Raihan Ramadhan
NPM : 22081010151
Degree Program : Bachelor (S1)
Study Program : Informatics
Faculty : Faculty of Computer Science

Hereby declares that this undergraduate thesis contains no part of any other scientific work that has been submitted to obtain an academic degree at any higher education institution. Furthermore, it does not contain any work or opinions previously written or published by others, except for those which are explicitly cited in this thesis and listed completely in references.

And I declare that this scientific document is free from elements of plagiarism. If in the future indications of plagiarism are found in this Thesis, I am willing to accept sanctions in accordance with the applicable laws and regulations.

Thus, I made this statement without any coercion from anyone and to be used as it should.



Surabaya, September 9 2026
Declarant,



RAIHAN RAMADHAN
NPM. 22081010151

ABSTRACT

Student Name / NPM: Raihan Ramadhan/22081010151
Thesis Title: Application of the CatBoost Algorithm with Optuna Optimization for Classifying the Etiology of Acute Fever Using SMOTE-NC as a Data Class Balancer
Supervisor: 1. Yisti Vita Via, S.ST., M.Kom.
2. Ardhon Rakhmadi, S.Tr.T., M.Kom.

This research addresses class imbalance and symptomatic overlap in the multiclass classification of acute fever etiology by integrating CatBoost and SMOTE-NC with hyperparameter optimization via Optuna, implemented as a mobile-based decision support system. The secondary dataset, derived from the AFIRE Study (2013–2016), yielded 801 clean samples across four target classes: Dengue, *Rickettsia spp.*, *Other bacterial*, and *Other viral*. Using Stratified 5-Fold Cross Validation, the optimal CatBoost+SMOTE-NC+Optuna model achieved an average accuracy of 74.16%, Macro Recall of 64.74%, and Macro F1-Score of 64.08%, with the strongest performance on Dengue (F1=0.87) and more moderate results for the minority *Rickettsia spp.* class (F1=0.51). SMOTE-NC significantly improved Macro Recall without a statistically significant loss in Accuracy. Statistical testing confirmed CatBoost significantly outperformed XGBoost across all metrics ($p < 0.05$), while showing comparable performance to Random Forest ($p > 0.05$); CatBoost was ultimately selected for its architectural efficiency and inference time of just 0.024 ms. A Brier Score-based calibration analysis further confirmed the reliability of the predicted probabilities. The final model was deployed in a client-server architecture (FastAPI backend, Flutter frontend) and successfully passed black-box testing.

Keywords: CatBoost, SMOTE-NC, Acute Fever, Multiclass Classification, Mobile Application, FastAPI, Flutter

ACKNOWLEDGEMENTS

All praise be to Allah SWT for all His blessings, guidance, and grace upon the author, so that this undergraduate thesis titled **“APPLICATION OF THE CATBOOST ALGORITHM WITH OPTUNA OPTIMIZATION FOR CLASSIFYING THE ETIOLOGY OF ACUTE FEVER USING SMOTE-NC AS A DATA CLASS BALANCER”** could be completed properly.

The author would like to express gratitude to Mrs. Yisti Vita Via, S.ST., M.Kom. as Supervisor I and Mr. Ardhon Rakhmadi, S.Tr.T., M.Kom. as Supervisor II, who were willing to take the time to provide guidance, suggestions, and motivation to the author. In addition, during the preparation of the thesis proposal, the author also received a great deal of assistance from various parties. For this, the author would like to thank:

1. Mrs. Prof. Dr. Ir. Novirina Hendrasarie, M.T. as Dean of the Faculty of Computer Science, Universitas Pembangunan Nasional “Veteran” Jawa Timur.
2. Mrs. Dr. Intan Yuniar Purbasari, S.Kom. MSc. as Coordinator of the Informatics Study Program, Faculty of Computer Science, Universitas Pembangunan Nasional Jawa Timur.
3. Mrs. Yisti Vita Via, S.ST. M.Kom., and Mr. Ardhon Rakhmadi, S.Tr.T., M.Kom., as Supervisors who patiently provided direction, guidance, input, and motivation during the process of preparing this thesis.
4. Mrs. Fetty Tri Anggraeny, S.Kom., M.Kom. and Mrs. Eva Yulia Puspaningrum, S.Kom., M.Kom. as the Examining Lecturers who provided constructive criticism, input, and corrections toward the improvement of this thesis.
5. The lecturers of the Informatics Study Program, Universitas Pembangunan Nasional Veteran Jawa Timur, for the guidance provided throughout the author's education.
6. The author's beloved mother, who has always played an important role throughout the author's life journey, always providing support, both material and moral. Always present with her best prayers, becoming a source of strength that never dims, so that the author could persevere to this day.

7. The author's beloved father, who taught the author to be a responsible person, to work hard, and not to give up easily, and whose guidance has always been the author's compass in life.
8. The author's beloved sibling, who has always provided support and input in successfully completing this period of study.
9. My beloved late aunt, Sucik Samiati, and her family, who offered encouragement, welcomed me, and supported me throughout my university studies.
10. dr. Galih Indra Purlistyarini, as the clinical validator who generously provided her time and clinical expertise to validate the research instrument and offer valuable insights from a medical practitioner's perspective.
11. My friends and classmates in the Informatics Study Program, who have struggled together throughout this academic journey, for their companionship, mutual support, and shared perseverance in facing every challenge along the way.
12. My friends at RUMPIGENG, who have also given me support and encouragement and stood by me during both good times and hard times.

The author realizes that there are many shortcomings in the preparation of this thesis. Therefore, constructive criticism and suggestions from all parties are greatly expected for the improvement of this thesis. Finally, with all of the author's limitations, it is hoped that this report can be useful for all parties in general and for the author in particular.

Surabaya, September 3 2026

Raihan Ramadhan

TABLE OF CONTENTS

APPROVAL SHEET	ii
APPROVAL SHEET	iii
STATEMENT OF ORIGINALITY	vii
ABSTRACT	viii
ACKNOWLEDGEMENTS	ix
TABLE OF CONTENTS	xi
LIST OF FIGURES	xvi
LIST OF TABLES	xix
LIST OF CODES	xxi
CHAPTER I INTRODUCTION	1
1.1 Background	1
1.2 Problem Formulation	3
1.3 Research Objectives	3
1.4 Research Benefits.....	4
1.5 Research Limitations	4
CHAPTER II LITERATURE REVIEW	6
2.1 Previous Research	6
2.2 Acute Fever and Clinical and Laboratory Parameters	10
2.2.1 Etiology of Acute Fever	10
2.2.2 Clinical and Laboratory Parameters.....	11
2.3 Machine Learning	12
2.4 Data Preprocessing.....	13
2.5 Data Balancing Algorithms.....	16
2.6 Classification.....	20

2.7	Categorical Boosting (CatBoost) Algorithm	21
2.8	Integration of SMOTE-NC and CatBoost	24
2.9	Model Performance Evaluation Method.....	25
2.10	Hyperparameter Tuning and Feature Importance.....	29
2.11	Statistical Significance Testing (Paired t-test).....	31
2.12	Probability Calibration Analysis	32
2.13	Application Design	33
2.14	Client-Server Architecture and REST API.....	33
2.15	Flutter Framework for Mobile Applications	35
2.16	FastAPI for Machine Learning Model Deployment.....	36
2.17	Expert Validation Instrument	37
CHAPTER III SYSTEM DESIGN AND IMPLEMENTATION		39
3.1	Research Stages	39
3.2	Hardware and Software Requirements	40
3.3	Literature Study	41
3.4	Data Source and Collection	41
3.5	Data Preprocessing	47
3.5.1	Data Filtering and Label Selection	47
3.5.2	Data Split (Training and Testing Data Division).....	50
3.5.3	Missing Value Handling.....	51
3.5.4	Categorical Data Encoding.....	51
3.5.5	Data Normalization (Min-Max Scaler).....	52
3.6	Application of SMOTE-NC.....	53
3.6.1	Identification of the Minority Class.....	55
3.6.2	Distance Calculation (K-Nearest Neighbors)	55
3.6.3	Synthesis of New Data	56

3.7	CatBoost Modeling	57
3.7.1	Oblivious Decision Tree Architecture	58
3.7.2	Multiclass Loss Function	60
3.7.3	Ordered Boosting (Chained Tree Formation)	60
3.7.4	Final Prediction Classification	62
3.8	Integration of the SMOTE-NC Algorithm and CatBoost Modeling.....	63
3.9	Model Evaluation.....	64
3.9.1	Multiclass Confusion Matrix	64
3.9.2	Calculation of Evaluation Metrics	65
3.10	System Architecture and Mobile Application Design	67
3.10.1	Back-End Development	67
3.10.2	Front-End Development.....	68
3.11	Testing Scenario.....	70
3.11.1	Hyperparameter Tuning Scenario	70
3.11.2	Comparative Model Testing	71
3.11.3	Statistical Significance Testing Scenario (Paired t-test).....	73
3.11.4	Probability Calibration Analysis Scenario	74
3.11.5	Mobile Application Testing (Black-Box Testing)	75
3.11.6	Expert Validation Scenario	75
3.12	Model Performance and System Functionality Analysis	76
3.12.1	Machine Learning Model Performance Analysis	76
3.12.2	Results Visualization	77
	CHAPTER IV RESULTS AND DISCUSSION	78
4.1	System Implementation	78
4.1.1	Data Acquisition	78
4.1.2	Preprocessing	80

4.1.3	Implementation of SMOTE-NC as a Data Balancer	89
4.1.4	Implementation of CatBoost with SMOTE-NC	90
4.1.5	Implementation of CatBoost with Optuna Optimization.....	93
4.1.6	Implementation of CatBoost with SMOTE-NC and Optuna.....	96
4.1.7	Implementation of XGBoost with SMOTE-NC and Optuna	106
4.1.8	Implementation of Random Forest with SMOTE-NC and Optuna.....	108
4.2	System Performance Testing	112
4.2.1	Performance Comparison of CatBoost Using SMOTE-NC With and Without Optuna	112
4.2.2	Performance Comparison of CatBoost Using Optuna With and Without SMOTE-NC.....	115
4.2.3	Performance Comparison of CatBoost with Random Forest and XGBoost.....	118
4.2.4	Clinical Feature Importance Analysis	123
4.2.5	Probability Calibration Analysis	124
4.3	Model Implementation Into the Application	126
4.3.1	Model Export.....	126
4.3.2	Model Implementation on FastAPI	127
4.3.3	Mobile Application Interface Implementation	130
4.4	System Functionality Testing (Black-Box Testing)	135
4.5	Expert (Doctor) Validation Results	139
CHAPTER V CONCLUSION.....		142
5.1	Conclusion.....	142
5.2	Suggestions.....	143
REFERENCES		144
APPENDIX.....		152
Appendix 1. Expert (Doctor) Interview Guide and Results		152

Appendix 2. Expert (Doctor) Validation Sheet.....	156
---	-----

LIST OF FIGURES

Figure 2.1 Standard Supervised Learning Process for Classification (Source: Bohani et al., 2021) [28].....	13
Figure 2. 2 Data Preprocessing Workflow (Source: ScienceDirect) [30]	14
Figure 2.3 A Comparison of Conventional Tree Architecture and Oblivious Tree Architecture in CatBoost	21
Figure 2.4 Basic REST API Communication Architecture (Source: Mitropoulos et al., 2021).....	34
Figure 2.5 Layered Architecture of the Flutter Framework (Source: Flutter Docs)	35
Figure 2.6 Asynchronous Data Processing Architecture in FastAPI.....	37
Figure 3.1 Research Stages.....	39
Figure 3.2 Data Acquisition Process	43
Figure 3.3 Data Preprocessing.....	47
Figure 3.4 Application of the SMOTE-NC Algorithm.....	54
Figure 3.5 CatBoost Algorithm Modeling.....	58
Figure 3.6 Illustration of a Multiclass Confusion Matrix	64
Figure 3.7 Mobile Application Design Wireframe.....	69
Figure 4.1 Raw Dataset Excerpt of Acute Fever Medical Records in Excel Format	79
Figure 4.2 Target Class Distribution and Data Snippet After Filtering.....	81
Figure 4.3 Summary Log of Data Distribution After Stratified K-Fold Splitting.....	83
Figure 4.4 Comparison of Results Before and After Missing Value Imputation ..	84
Figure 4.5 Comparison of the Target Column Before and After Encoding	86
Figure 4.6 Results of the Target Column Encoding Process.....	86
Figure 4.7 Comparison of Columns Before and After Normalization	88
Figure 4.8 Resulting Data After All Pre-processing Stages Are Complete.....	88
Figure 4.9 Comparison of Target Class Distribution Before and After SMOTE-NC	90
Figure 4.10 Execution Log for CatBoost Model Training with SMOTE-NC.....	92

Figure 4.11 Visualization of the Global Performance of the Baseline CatBoost + SMOTE-NC Model.....	93
Figure 4.12 Snippet of the Optuna Search Execution Log and Results.....	95
Figure 4.13 Search Space Boundaries and Best Hyperparameter Results Based on Optuna.....	98
Figure 4.14 Combined 5-Fold MultiClass Loss Decline Curve for the CatBoost Model.....	100
Figure 4.15 Summary of CatBoost Evaluation Using Optuna.....	101
Figure 4.16 Average Performance of the Optimal CatBoost Model per Etiology Class (All Folds)	102
Figure 4.17 Combined Confusion Matrix of the Optimal CatBoost Model	104
Figure 4.18 Execution Results of the XGBoost Model Classification Report ...	107
Figure 4.19 Execution Results of the Random Forest Model Classification Report	111
Figure 4.20 Comparison of Global Performance Between Default-Parameter CatBoost and CatBoost with Optuna	113
Figure 4.21 Results of the Statistical Significance Test (Paired t-test) for Optuna's Model Performance.....	114
Figure 4.22 Comparison of Global CatBoost Performance in the SMOTE-NC Study	115
Figure 4.23 Results of the SMOTE-NC Statistical Significance Test on Global Performance	117
Figure 4.24 Results of the Comparative Algorithm Benchmarking Study	119
Figure 4.25 Results of the Statistical Significance Test (Paired t-test) for Algorithm Comparison	120
Figure 4.26 Comparative Analysis of Training Time and Inference Time Between Algorithms	122
Figure 4.27 Feature Importance Chart for the Optimal CatBoost Model	123
Figure 4.28 Global Brier Score Evaluation Results for the CatBoost Model (5-Fold)	125
Figure 4.29 Clinical Parameter Payload Testing Interface on Swagger UI	129
Figure 4.30 CatBoost Prediction Output Response in JSON Format	130

Figure 4.31 Clinical Parameter Input Form Page Interface.....	131
Figure 4.32 Visualization of the Final Prediction Result and Probability Distribution.....	134
Figure 4.33 Application Interface Response During Empty Attribute Test Scenario	137
Figure 4.34 Application Interface Response During the Data Type Anomaly Test Scenario	138

LIST OF TABLES

Table 2.1 Previous Research	6
Table 3.1 Details of Variables in the Acute Fever Dataset	44
Table 3.2 Overview of the Acute Fever Patient Clinical and Laboratory Dataset	45
Table 3.3 Target Class Distribution in the Initial Dataset	46
Table 3.4 Target Data Filtering Simulation	49
Table 3.5 Events Per Variable (EPV) Values per Etiology Class	50
Table 3.6 Stratified 5-Fold Cross Validation Iteration Scenario	51
Table 3.7 Target Class Mapping	52
Table 3.8 Range of Numerical Feature Values	52
Table 3.9 Simulated Data After Normalization	53
Table 3.10 Clean Dataset Class Distribution (Before Split and SMOTE-NC)	55
Table 3.11 Simulation of SMOTE-NC Data Synthesis Results	57
Table 3.12 Summary of Leaf Index Mapping Results (Oblivious Tree)	60
Table 3.13 Summary of the Ordered Boosting Mechanism Iterations	61
Table 3.14 Summary of Final Prediction Probabilities (Softmax)	63
Table 3.15 Confusion Matrix Simulation	65
Table 3.16 Summary of the Simulated Classification Report	67
Table 3.17 Hyperparameter Tuning Scenario (Optuna Search Space)	70
Table 3.18 SMOTE-NC and Optuna Testing Scenario	72
Table 3.19 Algorithm Comparison Testing Scenario	73
Table 4.1 Evaluation Results of Baseline CatBoost Model on Balanced Data (5-Fold)	92
Table 4.2 Evaluation Results of the CatBoost Model with Optuna (Without SMOTE-NC)	96
Table 4.3 Optimal Hyperparameter Configuration from the Optuna Search	99
Table 4.4 Details of the CatBoost Model Evaluation Results with Optuna Optimization per Fold	102
Table 4.5 Details of the Average Performance of the Optimal CatBoost Model per Etiology Class	104
Table 4.6 Details of the XGBoost Model Evaluation Results per Fold	108

Table 4.7 Details of the Random Forest Model Evaluation Results per Fold	112
Table 4.8 Details of the Performance Comparison Between Default CatBoost and Optuna Optimization Results.....	114
Table 4.9 Details of CatBoost Performance Comparison Without and With SMOTE-NC (Optuna)	116
Table 4.10 Details of the Global Performance Comparison Between Algorithms	120
Table 4. 11 Details of the Training Time and Inference Latency Comparison Between Algorithms	122
Table 4.12 Results of Interface and Server Functionality Testing (Black-Box Testing).....	136
Table 4.13 Recapitulation of the Expert Validation Sheet Assessment Results .	140

LIST OF CODES

Program Code 4.1: Dimension Reduction and Target Selection	81
Program Code 4.2: Subset Splitting Using Stratified K-Fold.....	82
Program Code 4.3: Missing Value Imputation Using the Median.....	84
Program Code 4.4: Target Column Encoding.....	85
Program Code 4.5: Feature Normalization Using MinMaxScaler.....	87
Program Code 4.6: Implementation of Data Augmentation Using SMOTE-NC .	89
Program Code 4.7: Implementation of CatBoost with SMOTE-NC	91
Program Code 4.8: Objective Function for CatBoost Parameter Search	94
Program Code 4.9: Optuna Execution and CatBoost Model Training	95
Program Code 4.10: Objective Function for CatBoost Parameter Search	97
Program Code 4.11: Optuna Execution and Final Model Training	97
Program Code 4.12: Objective Function for XGBoost Parameter Search.....	106
Program Code 4.13: XGBoost Training and Evaluation Execution	106
Program Code 4.14: Categorical Data Transformation with Ordinal Encoder...	109
Program Code 4.15: Objective Function for Random Forest Parameter Search	109
Program Code 4.16: Random Forest Training and Evaluation Execution.....	110
Program Code 4.17: Model Serialization and Pre-processing Parameters	127
Program Code 4.18: Prediction Request Handling Using FastAPI	128
Program Code 4.19: Clinical Parameter Input Widget Declaration in Flutter....	132
Program Code 4.20: JSON Payload Sending Method to the Server.....	133