UNDERGRADUATE THESIS

UTILIZATION OF COMPLETE BLOOD COUNT FOR ANEMIA CLASSIFICATION MODELING USING CATBOOST AND SHAP INTERPRETABILITY ANALYSIS

VIKA RAFI ANA
NPM 22081010009

THESIS ADVISORS
Eva Yulia Puspaningrum, S.Kom., M.Kom.
Retno Mumpuni, S.Kom., M.Sc.

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

APPROVAL SHEET

UTILIZATION OF COMPLETE BLOOD COUNT FOR ANEMIA CLASSIFICATION MODELING USING CATBOOST AND SHAP INTERPRETABILITY ANALYSIS

By :
VIKA RAFI ANA
NPM. 22081010009

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 June 10, 2026:

Approved,

Eva Yulia Puspaningrum S.Kom., M.Kom. **NIP. 19890705 202121 2 002**	....	**(Advisor I)**
Retno Mumpuni S.Kom., M.Sc. **NIP. 19870716 202521 2 045**	....	**(Advisor II)**
Dr. Intan Yuniar Purbasari, S.Kom., M.Sc. **NIP. 19800602 202521 2 029**	....	**(Head Assessor)**
Hazna At Thooriqoh, S.Tr.Kom., M.Kom. **NIP. 19970320 202406 2 003**	....	**(Assessor I)**

Acknowledged by,

Dean of the Faculty of Computer Science

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

APPROVAL SHEET

UTILIZATION OF COMPLETE BLOOD COUNT FOR ANEMIA CLASSIFICATION MODELING USING CATBOOST AND SHAP INTERPRETABILITY ANALYSIS

By :

VIKA RAFI ANA

NPM. 22081010009

Approved to proceed to the Thesis Examination

Approved by,

Coordinator of the Informatics Study Program

Faculty of Computer Science

Dr. Intan Yuniar Purbasari, S.Kom., M.Sc.

NIP. 19800602 202521 2 029

STATEMENT OF ORIGINALITY

I am the undersigned:

Student Name	:	Vika Rafi Ana
NPM	:	22081010009
Degree Program	:	Bachelor (S1)
Study Program	:	Informatics
Faculty	:	Faculty of Computer Science

Here by 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 the 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 undergraduate thesis, I am willing to accept sanctions in accordance with the applicable laws and regulations.

Thus, I make this statement without any coercion from anyone and for proper use as intended.

Surabaya, July 10th, 2026

Declarant,

METERAI TEMPEL

VIKA RAFI ANA

NPM. 22081010009

ABSTRACT

Student Name / NPM	:	Vika Rafi Ana / 22081010009
Thesis Title	:	Utilization of Complete Blood Count for Anemia Classification Modeling Using CatBoost and SHAP Interpretability Analysis
Advisor	:	1. Eva Yulia Puspaningrum, S.Kom., M.Kom 2. Retno Mumpuni S.Kom., M.Sc

This study proposes an anemia classification model based on Complete Blood Count (CBC) data, as hematological parameters exhibit distinct patterns across different diagnosis classes. The input variables consist of CBC parameters, including HGB, RBC, MCV, MCH, MCHC, WBC, and PLT. The classification process was performed using the CatBoost algorithm, while SHAP was employed to explain the contribution of each CBC feature to the model predictions. The dataset utilized in this research was the *Anemia Types Classification* dataset obtained from Kaggle. Initially, the dataset comprised 1,281 samples with 14 CBC attributes and 9 diagnosis classes. Following the removal of duplicate records, the final dataset contained 1,232 samples. The cleaned dataset was partitioned into training and testing subsets using an 80:20 stratified split to preserve the class distribution. Three experimental settings were investigated, consisting of the baseline CatBoost model, CatBoost with class weighting, and CatBoost with both class weighting and hyperparameter tuning. Among these configurations, the CatBoost model with class weighting achieved the highest performance, obtaining Accuracy, Macro F1-score, and Weighted F1-score values of 1.0000 on the testing dataset. In contrast, applying hyperparameter tuning did not improve the predictive performance and instead resulted in a considerably longer training time. The SHAP analysis further revealed that HGB, MCV, MCH, MCHC, WBC, PLT, and RBC were the most influential CBC features contributing to the classification process. Overall, the findings demonstrate that incorporating class weighting enhanced the classification capability of CatBoost, whereas SHAP improved the interpretability

of the model by explaining how individual CBC features influenced the prediction outcomes.

Keywords: Anemia, Complete Blood Count, CatBoost, Class Weight, SHAP.

ACKNOWLEDGEMENTS

Above all, the author would like to thank God Almighty for His endless mercy, guidance, and blessings, which have accompanied every stage of this undergraduate thesis entitled **"Utilizing Complete Blood Count for Anemia Prediction Modeling Using CatBoost and SHAP Interpretability Analysis".** This work was completed in partial fulfillment of the requirements for the Bachelor's Degree in Computer Science at the Informatics Study Program, Faculty of Computer Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur.

The author is deeply grateful to everyone whose guidance, encouragement, prayers, and unwavering support contributed to the completion of this thesis. Therefore, sincere appreciation is extended to the following individuals:

1. **Mama and Papa**, who have always showered me with endless prayers, love, support, encouragement, and sacrifices, as well as my entire family, especially 만다, for always being there for me so I could complete my studies smoothly and on time.
2. **Eva Yulia Puspaningrum, S.Kom., M.Kom.**, as the First Advisor, and **Retno Mumpuni, S.Kom., M.Sc., M.T.**, as the Second Advisor, for generously dedicating their time, energy, and thoughts to provide guidance, advice, and valuable feedback from the very beginning until this thesis was completed smoothly without making the process unnecessarily difficult.
3. **Prof. Dr. Ir. Novirina Hendrasarie, M.T.**, Dean of the Faculty of Computer Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur.
4. **Dr. Intan Yuniar Purbasari, S.Kom., M.Sc.**, Coordinator of the Informatics Study Program, Faculty of Computer Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur.
5. YT22, GR63, Bang Windah, as well as Jummo and Ryul, who unknowingly became my biggest mood boosters throughout this chaotic thesis journey, even though I was stressed because Yuki was not in the grid in 2026 but ya.. thank you for all the funny moments that always managed to bring my mood back after spending the whole day dealing with all kinds of thesis stuff. And thank you to my online friends for being the safe place for myalteregogw.

6. My fellow thesis fighters who kept supporting, helping, and reminding each other throughout this journey. Even though we were all equally lost sometimes, we somehow made it through together. Tengkyu so much I love you all.
7. Haa... well well well for last but not least **thank you to myself**. Thank you for holding on for always trying for getting back up every time things felt overwhelming and for never giving up even though this thesis completely messed up my sleep schedule and stole time from my hobbies yeah writing this thesis really felt like riding a roller coaster but in the end I made it through everything. See? I did it.

The author realizes that this thesis still has many shortcomings after all, I'm only human and nobody's perfect. Therefore, any constructive criticism and suggestions from everyone are greatly appreciated to improve this thesis. Finally, despite all the limitations the author has, I hope this thesis will be beneficial to everyone.

Surabaya, July 10th 2026

Author

TABLE OF CONTENTS

UNDERGRADUATE THESIS .... **i**

APPROVAL SHEET .... **iii**

APPROVAL SHEET .... **v**

STATEMENT OF ORIGINALITY .... **vii**

ABSTRACT .... **ix**

ACKNOWLEDGEMENTS .... **xi**

TABLE OF CONTENTS .... **xiii**

LIST OF FIGURE .... **xvii**

LIST OF TABLE .... **xxi**

CHAPTER I INTRODUCTION **1**

1.1. Background 1

1.2. Research Questions 3

1.3. Research Objectives 3

1.4. Research Benefits 4

1.5. Research Scope 4

CHAPTER II LITERATURE REVIEW **5**

2.1. Previous Studies 5

2.2. Anemia 6

2.3. Complete Blood Count (CBC) 7

2.4. Machine Learning 8

2.5. Supervised Learning 9

2.6. Classification 11

2.7. CatBoost 12

2.7.1. Gradient Boosting Mechanism 13

2.7.2. Ordered Target Statistics (OTS) 14

2.7.3. Ordered Boosting 14
2.7.4. Oblivious Decision Tree 15
2.7.5. CatBoost Objective Function 15
2.7.6. Softmax Cross-Entropy Loss for Multiclass Classification 16
2.7.7. Leaf Values 18
2.7.8. Class Weights 19
2.8. Confusion Matrix 19
2.9. SHapley Additive exPlanations (SHAP) 21
CHAPTER III RESEARCH METHODOLOGY 23
3.1. Data Input 23
3.2. Data Acquisition 24
3.3. Data Preprocessing 25
3.3.1. Data Cleaning 25
3.3.2. Feature and Target Split 27
3.3.3. Train–Test Split 27
3.4. Class Weighting 28
3.5. CatBoost Model Development 31
3.6. Model Evaluation 36
3.7. Experimental Design 37
3.8. Model Interpretability Analysis 39
3.9. Model Implementation Design 45
CHAPTER IV RESULTS AND DISCUSSION 49
4.1. Data Preprocessing 49
4.1.1. Dataset Acquisition 49
4.1.2. Duplicate Data Removal 50
4.1.3. Missing Value Inspection 52

4.1.4. Feature and Target Separation 53
4.1.5. Training–Testing Data Split 54
4.2. Classification Model Development 56
4.2.1. Baseline CatBoost Model 56
4.2.2. CatBoost Model with Class Weights 60
4.2.3. CatBoost Model with Class Weights and Hyperparameter Tuning 63
4.3. Model Evaluation 67
4.3.1. Evaluation of the Baseline CatBoost Model 68
4.3.2. Evaluation of the CatBoost Model with Class Weights 70
4.3.3. Evaluation of the Tuned CatBoost Model 72
4.4. Comparison of Results Across Experimental Scenarios 74
4.5. Model Interpretability Analysis Using SHAP 81
4.6. Results of the System Interface Implementation 99
4.6.1. Results of the Home Page Implementation 100
4.6.2. Results of the CBC Data Input Page Implementation 100
4.6.3. Results of the Prediction Results Page Implementation 103
CHAPTER V CONCLUSIONS AND RECOMMENDATIONS 107
5.1. Conclusions 107
5.2. Recommendations 108
REFERENCE 109
ATTACHMENT 115

LIST OF FIGURE

Figure 2.1 Normal Red Blood Cells and Anemia [16] 7
Figure 2.2 Machine Learning Workflow [23] 9
Figure 2.3 Supervised Learning [27] 10
Figure 2.4 Classification Based on the Number of Classes [31] 11
Figure 2.5 CatBoost Flowchart [34] 12
Figure 2.6 Confusion Matrix [37] 20
Figure 3.1 Research Workflow 23
Figure 3.2 Training and Testing Data Split 28
Figure 3.3 Class Distributions of Y_train and Y_test 28
Figure 3.4 Class Weighting Workflow in the Model Training Process 28
Figure 3.5 CatBoost Model Development Workflow 31
Figure 3.6 Workflow of SHAP-Based Model Interpretability Analysis 39
Figure 3.7 Example of the SHAP Result for the Healthy Class 44
Figure 3.8 Mockup of the System Home Page 47
Figure 3.9 Mockup of the CBC Data Input Page 48
Figure 3.10 Mockup of the Prediction Results Page 48
Figure 4.1 Initial Dataset Information 50
Figure 4.2 Example of Duplicate Records to Be Removed 51
Figure 4.3 Duplicate Data Removal Results 52
Figure 4.4 Missing Value Inspection Results 53
Figure 4.5 Feature and Target Separation Results 54
Figure 4.6 Training and Testing Data Split Result 55
Figure 4.7 Detailed Distribution of the Training and Testing Sets 56
Figure 4.8 Complete Visualization of an Actual Decision Tree in the Baseline CatBoost Model 58
Figure 4.9 Enlarged View of the Upper Section of an Actual Decision Tree in the Baseline CatBoost Model 59
Figure 4.10 Enlarged View of an Actual Decision Tree Branch to the Leaf Node in the Baseline CatBoost Model 59
Figure 4.11 Class Weight Computation Results 61

Figure 4.12 5-Fold Cross-Validation Result for the Best Parameter Configuration 64
Figure 4.13 Hyperparameter Tuning Results 66
Figure 4.14 Evaluation Results of the Baseline CatBoost Model 68
Figure 4.15 Confusion Matrix of the Baseline CatBoost Model 69
Figure 4.16 Evaluation Results of the CatBoost Model with Class Weights 71
Figure 4.17 Confusion Matrix of the CatBoost Model with Class Weights 71
Figure 4.18 Evaluation of the Tuned CatBoost Model 73
Figure 4.19 Evaluation of the Tuned CatBoost Model 74
Figure 4.20 Training Data Distribution 75
Figure 4.21 Class Weight Values for Each Diagnosis Class 76
Figure 4.22 Prediction Audit of the CatBoost Model with Class Weights 77
Figure 4.23 Overfitting Assessment of the CatBoost Model with Class Weights 78
Figure 4.24 Summary of Results for Scenarios 1 and 2 79
Figure 4.25 Comparison Between Cross-Validation and Testing Performance ... 79
Figure 4.26 SHAP Beeswarm Plot Visualizations for All Diagnosis Classes Healthy 84
Figure 4.27 SHAP Beeswarm Plot for the Healthy Class 85
Figure 4.28 SHAP Beeswarm Plot for the Iron Deficiency Anemia Class 87
Figure 4.29 SHAP Beeswarm Plot for the Leukemia Class 89
Figure 4.30 SHAP Beeswarm Plot for the Leukemia with Thrombocytopenia Class 90
Figure 4.31 SHAP Beeswarm Plot for the Macrocytic Anemia Class 91
Figure 4.32 SHAP Beeswarm Plot for the Normocytic Hypochromic Anemia Class 93
Figure 4.33 SHAP Beeswarm Plot for the Normocytic Normochromic Anemia Class 94
Figure 4.34 SHAP Beeswarm Plot for the Other Microcytic Anemia Class 96
Figure 4.35 SHAP Beeswarm Plot for the Thrombocytopenia Class 98
Figure 4.36 Results of the Home Page Implementation 100
Figure 4.37 Results of the CBC Data Input Page Implementation 101
Figure 4.38 Example of CBC Data Entry in the System 101

Figure 4.39 Example of CBC data Entry in the System 102
Figure 4.40 Model Accuracy Visualization 103
Figure 4.41 Visualization of Feature Contributions to the Healthy Class Prediction 104
Figure 4.42 Visualization of Feature Contributions to the Normocytic Hypochromic Anemia Class Prediction 105

LIST OF TABLE

Table 3.1 Dataset Attributes 24
Table 3.2 Diagnosis Class Distribution 25
Table 3.3 Class Distribution After Duplicate Removal 26
Table 3.4 Feature and Target Split 27
Table 3.5 Number of Samples in Each Diagnosis Class 29
Table 3.6 Class Weights Based on the Sample Distribution of Each Diagnosis Class 30
Table 3.7 Example of Aggregated Gradient and Hessian Values 35
Table 3.8 Scenario 1 (CatBoost with Default Parameters: Baseline vs. Class Weights) 38
Table 3.9 Scenario 2 (Best Model from Scenario 1 with Hyperparameter Tuning) 39
Table 3.10 Example of CatBoost Logit Outputs 40
Table 3.11 Illustration of Feature Subsets 41
Table 3.12 Example of SHAP Values for the 14 CBC Features 42
Table 3.13 Reference Intervals and Characteristics of CBC Parameters for the Healthy Class 45
Table 4.1 Result of Scenario 1 with Two Sub-Conditions 75
Table 4.2 Results of Scenario 2 (CatBoost with Class Weights and Hyperparameter Tuning) 78
Table 4.3 Summary of Evaluation Results Across All Experimental Scenarios .. 80