



**UNDERGRADUATE THESIS**

**APPLICATION OF ENSEMBLE MACHINE  
LEARNING FOR PHISHING SITE  
IDENTIFICATION BASED ON URL AND VISUAL  
ANALYSIS**

**PASKALIS REYNALDY ELROY GABRIEL**  
NPM 22081010197

**THESIS ADVISORS**

Dr. Eng. Ir. Anggraini Puspita Sari, ST., MT.  
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

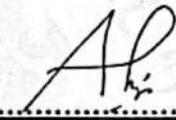
### APPLICATION OF ENSEMBLE MACHINE LEARNING FOR PHISHING SITE IDENTIFICATION BASED ON URL AND VISUAL ANALYSIS

By:  
PASKALIS REYNALDY ELROY GABRIEL  
NPM. 22081010197


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 April 17, 2026:

Approved,

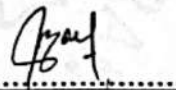
Dr. Eng. Ir. Anggraini Puspita Sari, ST., MT  
NPT. 222198 60 816400

  
..... (Advisor I)

Achmad Junaidi, S.Kom., M.Kom  
NIP. 197811102025211048

  
..... (Advisor II)

Made Hanindia Prami Swari, S.Kom, M.Cs  
NIP. 19890205 2018032 001


  
..... (Head Assessor)

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

  
..... (Assessor I)

Acknowledge by,

Dean of the Faculty of Computer Science

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

## APPROVAL SHEET

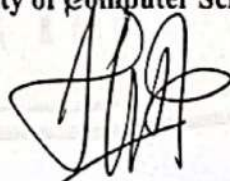
### APPLICATION OF ENSEMBLE MACHINE LEARNING FOR PHISHING SITE IDENTIFICATION BASED ON URL AND VISUAL ANALYSIS

By:  
PASKALIS REYNALDY ELROY GABRIEL  
NPM. 22081010197

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 : Paskalis Reynaldy Elroy Gabriel  
NPM : 22081010197  
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, 12 May 2026  
Declarant,



PASKALIS REYNALDY ELROY GABRIEL  
NPM. 22081010197

## ABSTRACT

Student Name / NPM : Paskalis Reynaldy Elroy Gabriel / 22081010197  
Thesis Title : Application Of Ensemble Machine Learning For Phishing Site Identification Based On URL And Visual Analysis  
Advisor : 1. Dr. Eng. Ir. Anggraini Puspita Sari, ST., MT.  
2. Achmad Junaidi, S.Kom., M.Kom.

Phishing attacks continue to evolve as a significant cybersecurity threat, with traditional blacklist-based and rule-based approaches proving insufficient in detecting newly emerging phishing sites. This study proposes an ensemble-based phishing detection system that integrates two analytical modalities: URL analysis using TF-IDF with character n-gram (3,6) and Complement Naïve Bayes (CNB), and visual web page analysis using VGG16 as a feature extractor and XGBoost as a classifier. Final classification decisions are produced through a late score fusion mechanism with weights of 0.8 for the URL pathway and 0.2 for the visual pathway. The system was evaluated using 8,707 URL samples and 824 website screenshot samples. Experimental results demonstrate that the hybrid ensemble system achieves an Accuracy of 94.90%, Precision of 0.9417, Recall of 0.9303, F1-Score of 0.9360, and ROC-AUC of 0.9801 — outperforming both the URL-only model (94.51%) and the visual-only model (86.29%). The 0.8/0.2 weight configuration was selected based on the visual pathway's ability to suppress false positives and provide an independent verification layer against URL obfuscation, consistent with the late fusion principle that minority modality contributions are most effective at weights  $\geq 0.15$ . This study demonstrates that a multimodal ensemble approach significantly enhances the robustness and accuracy of phishing detection compared to unimodal approaches.

**Keywords:** Phishing, Ensemble Learning, TF-IDF, Complement Naïve Bayes, VGG16, XGBoost, Late Score Fusion

## ACKNOWLEDGEMENTS

All praise and gratitude are rendered to the presence of God Almighty. for His abundant grace, guidance, and mercy, so that the author has been able to complete the preparation of this thesis entitled "**Application of Ensemble Machine Learning for Phishing Site Identification Based on URL And Visual Analysis**". The writing of this thesis is the final stage that must be completed to obtain the degree of Bachelor of Computer Science in the Informatics Study Program, Faculty of Computer Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur.

In the process of preparing this thesis, the author acknowledges that the achievements accomplished were inseparable from the assistance, support, and prayers of various parties who have contributed both morally, spiritually, and materially. Therefore, with all humility, the author extends sincere gratitude to:

1. Prof. Dr. Ir. Novirina Hendrasarie, M.T. as the Dean of the Faculty of Computer Science, National Development University "Veteran" East Java.
2. Dr. Intan Yuniar Purbasari, S.Kom. MSc. as Head of the Informatics Study Program.
3. Retno Mumpuni, S.Kom., M.Sc as the author's academic advisor who has provided assistance in academic counseling from the beginning of studies until the end.
4. Dr. Eng. Ir. Anggraini Puspita Sari, ST., MT. and Achmad Junaidi, S.Kom., M.Kom as Thesis Supervisors who have guided and provided direction so that the author was able to complete this thesis.
5. Made Hanindia Prami Swari, S.Kom, M.Cs and Fetty Tri Anggraeny, S.Kom. M.Kom as Thesis Examiners who have also directed the author so that this thesis could be properly prepared.
6. Andreas Nugroho Sihananto, S.Kom., M.Kom., who has assisted the author in completing all administrative matters related to the thesis defense proceedings through to graduation.
7. All lecturers of the Informatics Study Program at UPN "Veteran" Jawa Timur who have imparted knowledge, guidance, and inspiration

of great significance in the author's academic development

8. My dearest Mother and Father, whose endless support, prayers, and guidance have been the writer's greatest strength and foundation throughout this journey.
9. Stefanie Mareta Angeline who has always motivated the author to complete this thesis, offering companionship, support, and encouragement until the very end.
10. The author's closest circles and support systems, Bendul Santuy, Densus88, and Rindu Rumah, for the friendship, laughter, and solidarity that kept the author's spirits high during the most challenging times of this study.
11. All members and administrators of St. Patrick Catholic Student Family and the Informatics Student Association at UPN "Veteran" Jawa Timur, for the invaluable lessons in camaraderie, teamwork, and leadership.
12. All parties who cannot be mentioned one by one, who have been part of the author's journey and have grown together in the academic world.

The author realizes that in the preparation of the following thesis there are many shortcomings. Hence, constructive criticism and suggestions from all parties are highly expected for the perfection of writing the following thesis. With all the limitations that the author has, hopefully the following report can be useful for all parties in general and the author in particular.

Surabaya, May 6<sup>th</sup> 2026

Author

## TABLE OF CONTENTS

<b>APPROVAL SHEET .....</b>	<b>iii</b>
<b>APPROVAL SHEET .....</b>	<b>v</b>
<b>STATEMENT OF ORIGINALITY .....</b>	<b>vii</b>
<b>ACKNOWLEDGEMENTS.....</b>	<b>xi</b>
<b>TABLE OF CONTENTS.....</b>	<b>xiii</b>
<b>LIST OF FIGURES .....</b>	<b>xv</b>
<b>LIST OF TABLES .....</b>	<b>xviii</b>
<b>LIST OF CODE .....</b>	<b>xx</b>
<b>CHAPTER I INTRODUCTION.....</b>	<b>1</b>
1.1. Background .....	1
1.2. Research Questions .....	3
1.3. Research Objectives .....	3
1.4. Research Benefits.....	4
1.5. Scope of Research.....	5
<b>CHAPTER II LITERATURE REVIEW.....</b>	<b>6</b>
2.1. Previous Research .....	6
2.2. Phishing and Website Phishing.....	8
2.3. Term Frequency–Inverse Document Frequency (TF-IDF).....	11
2.4. Complement Naïve Bayes.....	14
2.5. Convolutional Neural Network (CNN).....	16
2.6. Extreme Gradient Boosting (XGBoost).....	19
2.7. Normalization.....	21
2.8. Hyperparameter Tuning .....	24
2.9. Imbalanced Data .....	27
2.10. Ensemble Method.....	31
2.11. Model Evaluation Metrics.....	33
2.12. Selenium WebDriver.....	36
<b>CHAPTER III METHODOLOGY .....</b>	<b>37</b>
3.1. Research Method.....	37
3.2. Requirement Analysis .....	38

3.3.	Data Collection .....	40
3.4.	Data Preprocessing.....	41
3.5.	Model Development.....	53
3.6.	Testing Scenarios .....	64
3.7.	Model Evaluation.....	65
3.8.	GUI Design .....	68
<b>CHAPTER IV RESULTS AND DISCUSSION .....</b>		<b>69</b>
4.1.	Data Preparation.....	69
4.2.	Data Pre-Processing .....	71
4.3.	Model Implementation.....	77
4.4.	Hyperparameter Testing Scenarios .....	83
4.5.	Performance Evaluation Results .....	144
4.6.	User Interface Implementation.....	154
<b>CHAPTER V CONCLUSION &amp; RECOMMENDATIONS.....</b>		<b>163</b>
5.1.	Conclusion .....	163
5.2.	Recommendations for Future Development .....	164
<b>BIBLIOGRAPHY .....</b>		<b>165</b>

## LIST OF FIGURES

Figure 2. 1 Example of a Phishing Website Mimicking Another Site.....	9
Figure 2. 2 Data on the Top 10 Scam Modalities from OJK .....	10
Figure 2. 3 How CNN Works .....	16
Figure 2. 4 How XGBoost Works.....	20
Figure 2. 5 Ilustration of SMOTE .....	29
Figure 3. 1 Research Flowchart .....	37
Figure 3. 2 Preprocessing Flowchart.....	42
Figure 3. 3 URL Path Preprocessing Flowchart .....	42
Figure 3. 4 Visual Path Preprocessing Flowchart .....	47
Figure 3. 5 5x5 Pixels Sampled from the Dataset.....	48
Figure 3. 6 URL Path Flowchart.....	54
Figure 3. 7 Visual Path Flowchart .....	59
Figure 3. 8 VGG16 Feature Extraction Flowchart.....	60
Figure 3. 9 Ensemble Method Flowchart.....	62
Figure 3. 10 GUI Design.....	68
Figure 4. 1 Sample Contents of the URL Dataset.....	69
Figure 4. 2 Sample Legitimate Image Dataset.....	70
Figure 4. 3 Sample Phishing Image Dataset .....	71
Figure 4. 4 Results for Threshold Value 0.3 .....	85
Figure 4. 5 Results for Threshold Value 0.4 .....	86
Figure 4. 6 Results for Threshold Value 0.5 .....	87
Figure 4. 7 Results for Threshold Value 0.6 .....	88
Figure 4. 8 Results for Threshold Value 0.7 .....	89
Figure 4. 9 Results for Link 0.5 / Visual 0.5.....	91
Figure 4. 10 Results for Link 0.6 / Visual 0.4.....	92
Figure 4. 11 Results for Link 0.7 / Visual 0.3.....	93
Figure 4. 12 Results for Link 0.8 / Visual 0.2.....	94
Figure 4. 13 Results for Link 0.9 / Visual 0.1.....	95
Figure 4. 14 Results for Link 1.0 / Visual 0.0.....	96
Figure 4. 15 Confusion Matrix Results for Alpha 0.1.....	98

Figure 4. 16 Performance Metrics for Alpha 0.1 .....	99
Figure 4. 17 Confusion Matrix Results for Alpha 0.5.....	100
Figure 4. 18 Performance Metrics for Alpha 0.5 .....	101
Figure 4. 19 Confusion Matrix Results for Alpha 1.0.....	102
Figure 4. 20 Performance Metrics for Alpha 1.0 .....	103
Figure 4. 21 Confusion Matrix Results for Alpha 2.0) .....	104
Figure 4. 22 Performance Metrics for Alpha 2.0 .....	105
Figure 4. 23 Confusion Matrix Results for Alpha 5.0.....	106
Figure 4. 24 Performance Metrics for Alpha 5.0 .....	107
Figure 4. 25 Confusion Matrix Results for N-gram (2,4) .....	109
Figure 4. 26 Performance Metrics for N-gram (2,4).....	110
Figure 4. 27 Confusion Matrix Results for N-gram (2,7) .....	111
Figure 4. 28 Performance Metrics for N-gram (2,7).....	112
Figure 4. 29 Confusion Matrix Results for N-gram (3,5) .....	113
Figure 4. 30 Performance Metrics for N-gram (3,5).....	114
Figure 4. 31 Confusion Matrix Results for N-gram (3,6) .....	115
Figure 4. 32 Performance Metrics for N-gram (3,6).....	116
Figure 4. 33 Confusion Matrix Results for N-gram (3,7) .....	117
Figure 4. 34 Performance Metrics for N-gram (3,7).....	118
Figure 4. 35 Confusion Matrix Results for N_estimators 50.....	120
Figure 4. 36 Performance Metrics for N_estimators 50.....	121
Figure 4. 37 Confusion Matrix Results for N_estimators 100.....	122
Figure 4. 38 Performance Metrics for N_estimators 100.....	123
Figure 4. 39 Confusion Matrix Results for N_estimators 200.....	124
Figure 4. 40 Performance Metrics for N_estimators 200.....	125
Figure 4. 41 Confusion Matrix Results for N_estimators 300.....	126
Figure 4. 42 Performance Metrics for N_estimators 300.....	127
Figure 4. 43 Confusion Matrix Results for N_estimators 500.....	128
Figure 4. 44 Performance Metrics for N_estimators 500.....	129
Figure 4. 45 Confusion Matrix Results for Epoch 5 .....	131
Figure 4. 46 CNN Training Loss and Accuracy Graph for Epoch 5 .....	132
Figure 4. 47 Ensemble Performance Metrics for Epoch 5 .....	132

Figure 4. 48 Confusion Matrix Results for Epoch 10 .....	133
Figure 4. 49 CNN Training Loss and Accuracy Graph for Epoch 10 .....	134
Figure 4. 50 Ensemble Performance Metrics for Epoch 10 .....	135
Figure 4. 51 Confusion Matrix Results for Epoch 15 .....	136
Figure 4. 52 CNN Training Loss and Accuracy Graph for Epoch 15 .....	137
Figure 4. 53 Ensemble Performance Metrics for Epoch 15 .....	138
Figure 4. 54 Confusion Matrix Results for Epoch 20 .....	139
Figure 4. 55 CNN Training Loss and Accuracy Graph for Epoch 20 .....	140
Figure 4. 56 Ensemble Performance Metrics for Epoch 20 .....	141
Figure 4. 57 Confusion Matrix Results for Epoch 30 .....	142
Figure 4. 58 CNN Training Loss and Accuracy Graph for Epoch 30 .....	142
Figure 4. 59 Ensemble Performance Metrics for Epoch 30 .....	143
Figure 4. 60 CNB Model Confusion Matrix .....	146
Figure 4. 61 CNB Model Classification Report.....	147
Figure 4. 62 VGG16 Model Confusion Matrix.....	148
Figure 4. 63 VGG16 Training History Graph.....	148
Figure 4. 64 VGG16 Model Classification Report .....	149
Figure 4. 65 XGBoost Model Confusion Matrix .....	150
Figure 4. 66 XGBoost Model Classification Report.....	151
Figure 4. 67 XGBoost Training Curve.....	152
Figure 4. 68 Main System Interface Display .....	157
Figure 4. 69 Interface Display When Screenshot Capture Fails .....	158
Figure 4. 70 Legitimate Website Result Display .....	159
Figure 4. 71 Grad-CAM Heatmap for Legitimate Website .....	160
Figure 4. 72 Phishing Website Result Display .....	161
Figure 4. 73 Phishing Website Result for Phishing Website .....	162

## LIST OF TABLES

Table 2. 1 Confusion Matrix .....	34
Table 3. 1 Hardware Requirements.....	38
Table 3. 2 Software Requirements .....	39
Table 3. 3 Initial URL Samples (Before Preprocessing).....	43
Table 3. 4 Results After Missing Values Removal .....	43
Table 3. 5 Results After Structure Validation.....	44
Table 3. 6 Results After Normalization .....	45
Table 3. 7 Results After Empty URL Check .....	45
Table 3. 8 Results After Duplication Removal .....	46
Table 3. 9 Final URL Preprocessing Results .....	46
Table 3. 10 Original Image Sample (5×5 Pixels, Red Channel).....	49
Table 3. 11 Resize Result (3×3 Pixels, R Channel).....	50
Table 3. 12 Pixel Normalization Result (3×3, Red Channel) .....	51
Table 3. 13 SMOTE Results .....	52
Table 3. 14 Training Set Statistics .....	52
Table 3. 15 StandardScaler Results.....	53
Table 3. 16 Example TF-IDF Vector for URL 1 .....	56
Table 3. 17 TF-IDF Vector After L2 Normalization (Input for CNB) .....	57
Table 3. 18 Training Data and Term Counts.....	57
Table 3. 19 XGBoost Prediction Results .....	61
Table 3. 20 Ensemble Results Across Various Scenarios.....	63
Table 3. 21 Hyperparameters Tested .....	64
Table 3. 22 Evaluation Dataset .....	65
Table 3. 23 Confusion Matrix .....	66
Table 4. 1 URL Dataset Distribution .....	74
Table 4. 2 Image Dataset Distribution .....	77
Table 4. 3 Summary of Threshold Scenario Testing Results.....	90
Table 4. 4 Summary of Fusion Weight Scenario Testing Results .....	97
Table 4. 5 Summary of CNB Alpha Scenario Testing Results .....	108
Table 4. 6 Summary of N-gram Range Scenario Testing Results .....	119

Table 4. 7 Summary of N_estimators Scenario Testing Results.....	130
Table 4. 8 Summary of Epoch Scenario Testing Results.....	144
Table 4. 9 Optimal Parameter Configuration.....	145
Table 4. 10 Individual vs Ensemble Model Performance Comparison .....	153
Table 4. 11 Model Performance Comparison by Pathway .....	153

## LIST OF CODE

Code Program 4. 1 URL Validation Implementation .....	72
Code Program 4. 2 Normalization Implementation .....	73
Code Program 4. 3 URL Dataset Cleaning Process .....	73
Code Program 4. 4 Image Validation Implementation .....	75
Code Program 4. 5 Image Preprocessing Implementation.....	76
Code Program 4. 6 TF-IDF Configuration and Vectorization .....	78
Code Program 4. 7 Data Splitting and Complement Naive Bayes Training.....	78
Code Program 4. 8 VGG16 Architecture Implementation.....	79
Code Program 4. 9 VGG16 Training .....	80
Code Program 4. 10 VGG16 Feature Extraction .....	81
Code Program 4. 11 SMOTE and StandardScaler .....	81
Code Program 4. 12 Training XGBoost.....	82
Code Program 4. 13 Ensemble Fusion Implementation.....	83
Code Program 4. 14 URL Pathway Analysis and Screenshot .....	155
Code Program 4. 15 Visual Pathway Analysis and Ensemble Fusion.....	156
Code Program 4. 16 Grad-CAM Visualization.....	156