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 ≥ 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 6th 2026

Author

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF CODE