UNDERGRADUATE THESIS

IMPLEMENTATION OF A HYBRID AUTOENCODER-CATBOOST MODEL FOR REAL-TIME PHISHING SITE DETECTION BASED ON URL AND HTML ANALYSIS

MUHAMMAD HILMY AZIZ
NPM 22081010324

THESIS ADVISORS
Dr. Rizky Parlika, S.Kom, M.Kom.
Dr. Ir. Kartini, S.Kom, MT.

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

APPROVAL SHEET

IMPLEMENTATION OF A HYBRID AUTOENCODER-CATBOOST MODEL FOR REAL-TIME PHISHING SITE DETECTION BASED ON URL AND HTML ANALYSIS

By:
MUHAMMAD HILMY AZIZ
NPM. 22081010324

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,

Dr. Rizky Parlika, S.Kom, M.Kom.
NIP. 19840518 202121 1 003 ..... **(Advisor I)**

Dr. Ir. Kartini, S.Kom, MT.
NIP. 19611110 199103 2 001 ..... **(Advisor II)**

Budi Nugroho, S.Kom, M.Kom.
NIP. 19800907 202121 1 005 ..... **(Head Assessor)**

Ardhon Rakhmadi, S.Tr.T., M.Kom.
NIP. 19910805 202406 1 002 ..... **(Assessor I)**

Acknowledge by,

Dean of the Faculty of Computer Science

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

APPROVAL SHEET

IMPLEMENTATION OF A HYBRID AUTOENCODER-CATBOOST MODEL FOR REAL-TIME PHISHING SITE DETECTION BASED ON URL AND HTML ANALYSIS

By:
MUHAMMAD HILMY AZIZ
NPM. 22081010324

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	:	Muhammad Hilmy Aziz
NPM	:	22081010324
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 1, 2026
Declarant,

MUHAMMAD HILMY AZIZ
NPM. 22081010324

ABSTRACT

Student Name / NPM	:	Muhammad Hilmy Aziz / 22081010324
Thesis Title	:	Implementation of a Hybrid Autoencoder-CatBoost Model for Real-Time Phishing Site Detection Based on URL and HTML Analysis
Supervisor	:	1. Dr. Rizky Parlika, S.Kom, M.Kom. 2. Dr. Ir. Kartini, S.Kom, MT.

Phishing is one of the dominant cybersecurity threats involving fraudulent attempts to obtain users' sensitive information by imitating the appearance and flow of legitimate services. Conventional detection methods such as blacklisting and whitelisting are static in nature and therefore struggle to recognize new (zero-day) phishing attacks. This research aims to build a hybrid Autoencoder-CatBoost model to detect phishing websites in real time based on URL and HTML feature analysis, to evaluate its performance using classification evaluation metrics, and to implement it as a browser extension. The Autoencoder is used to learn a latent representation from 49 features (31 HTML features and 18 URL features), which are subsequently classified using the CatBoost algorithm. Testing results show that the model configuration with a latent dimension of 36 and an 80:10:10 data split ratio provides the best balance between accuracy and computational efficiency, achieving an accuracy of 94.68%, an F1-score of 93.53%, and an inference time of 0.4445 ms, or about 2.2 times faster than the baseline CatBoost model. The selected model was successfully implemented as a Chromium-based browser extension integrated with a backend API, and all black-box testing scenarios were declared successful. These results indicate that the hybrid Autoencoder-CatBoost model can provide an accurate, adaptive, and efficient phishing detection solution suitable for real-time deployment.

Keywords: Phishing, Autoencoder, CatBoost, Real-Time Detection, Browser Extension

ACKNOWLEDGEMENTS

Praise and thanks be to Allah SWT for all His mercy, guidance, and blessings, which have enabled the author to successfully complete this thesis titled **"Implementation of a Hybrid Autoencoder-CatBoost Model for Real-Time Phishing Site Detection Based on URL and HTML Analysis"**.

The author would like to thank Mr. Dr. Rizky Parlika, S.Kom., M.Kom., as the primary advisor, and Ms. Dr. Ir. Kartini, S.Kom., MT., as the co-advisor, for taking the time to provide invaluable guidance, advice, and motivation to the author. The author also received a great deal of support from various parties, whether moral, spiritual, or material. For this, the author would like to thank:

1. My beloved parents and family who provided endless moral and material support, which always accompanied the author, gave invaluable strength at every step of this journey.
2. Prof. Dr. Ir. Novirina Hendrasarie, M.T., Dean of the Faculty of Computer Science at the Universitas Pembangunan Nasional Veteran Jawa Timur.
3. Dr. Intan Yuniar Purbasari, S.Kom., M.Sc., Coordinator of the Informatics Study Program at the Universitas Pembangunan Nasional Veteran Jawa Timur.
4. Mr. Budi Nugroho, S.Kom., M.Kom., and Mr. Ardhon Rakhmadi, S.Tr.T., M.Kom., as thesis examiners, for their advice, feedback, and constructive criticism
5. Liana Firla Rizkika, partner and also my support system, who always provided support and listened to my complaints from the beginning to the end of writing this thesis.
6. My friends from Trenggalek, Alvin, Bintang, Dalta, Fadhil, Faliq, Farhan, Ghani, Kimin, Ndaru, Ronald, Ryan, Yumna, Yuna, and many others whom the author cannot mention one by one.
7. My college friends, Ade Tuban, Akbar Probolinggo, Bagus Wiguna, Bayu Lamongan, Rajawali Probolinggo, and many others whom the author cannot mention one by one.
8. All other parties whom the author cannot mention one by one.

The author acknowledges that this thesis is far from perfect. Therefore, the author would greatly appreciate constructive criticism and suggestions. In closing, may this thesis contribute to the advancement of science, particularly in the field of computer science.

Surabaya, September 1, 2026

Author,

Muhammad Hilmy Aziz

NPM. 22081010324

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF PSEUDOCODES