



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 Parluka, 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:

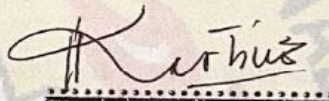
Approved,

Dr. Rizky Parluka, S.Kom, M.Kom.
NIP. 19840518 202121 1 003



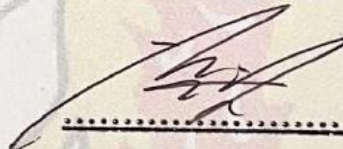
(Advisor I)

Dr. Ir. Kartini, S.Kom, MT.
NIP. 19611110 199103 2 001



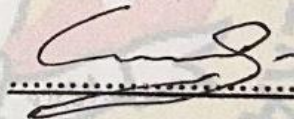
(Advisor II)

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

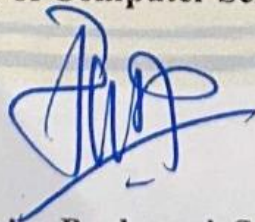
IMPLEMENTATION OF A HYBRID AUTOENCODER-CATBOOST MODEL FOR REAL-TIME PHISHING SITE DETECTION BASED ON URL AND HTML ANALYSIS

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

COVER PAGE	i
APPROVAL SHEET	iii
APPROVAL SHEET	v
STATEMENT OF ORIGINALITY	vi
ABSTRACT	viii
ACKNOWLEDGEMENTS.....	xi
TABLE OF CONTENTS.....	xiii
LIST OF FIGURES	xviii
LIST OF TABLES	xxi
LIST OF PSEUDOCODES	xxiii
CHAPTER I INTRODUCTION.....	1
1.1 Background	1
1.2 Research Questions	3
1.3 Research Objectives	3
1.4 Research Contributions	3
1.5 Scope and Limitations.....	4
CHAPTER II LITERATURE REVIEW.....	5
2.1 Previous Studies	5
2.2 Phishing.....	9
2.2.1 Definition of Phishing.....	9
2.2.2 Types and Media of Phishing.....	9
2.2.3 Characteristics of Phishing Website.....	9
2.2.4 Challenges in Phishing Detection Systems.....	10
2.3 Uniform Resource Locator (URL).....	10
2.3.1 Definition of URL	10
2.3.2 URL Structure and Components	11
2.4 Transfer Protocols	12
2.4.1 Hypertext Transfer Protocol (HTTP).....	12
2.4.2 Hypertext Transfer Protocol Secure (HTTPS).....	12
2.5 HyperText Markup Language (HTML)	12

2.5.1 Definition of HTML.....	12
2.5.2 Functions Of HTML	13
2.5.3 HTML Basic Structure.....	13
2.6 Browser Extensions.....	14
2.6.1 Definition of Browser Extensions	14
2.6.2 Function of Browser Extensions	14
2.6.3 Browser Extension Architecture	15
2.7 Machine Learning	15
2.7.1 Definition of Machine Learning.....	15
2.7.2 Types of Machine Learning	15
2.7.3 Stages of Machine Learning Implementation	17
2.7.4 Model Evaluation	18
2.8 Deep Learning.....	19
2.8.1 Definition of Deep Learning	19
2.8.2 Development of Deep Learning.....	20
2.9 Autoencoder	21
2.10 Gradient Boosting	25
2.10.1 XGBoost.....	26
2.10.2 LightGBM.....	27
2.10.3 CatBoost.....	28
CHAPTER III RESEARCH METHODOLOGY	33
3.1 Research Method.....	33
3.2 Data Acquisition.....	34
3.3 Data Preprocessing.....	36
3.3.1 Data Cleaning.....	37
3.3.2 Feature Extraction	37
3.3.3 Data Splitting	45
3.3.4 Data Normalization	45
3.4 Autoencoder	47
3.4.1 Encoder Process	48
3.4.2 Decoder Process	49
3.4.3 Error Calculation.....	50

3.4.4 Autoencoder Implementation.....	50
3.5 CatBoost.....	51
3.5.1 Numerical Features Quantization.....	51
3.5.2 Decision Tree Evaluation and Model Update	52
3.6 Evaluation	55
3.7 Testing Scenarios	57
3.7.1 Data Split Proportion	57
3.7.2 Architecture Comparison	57
3.7.3 Latent Dimension Variation.....	58
3.7.4 Model Inference Time.....	59
3.7.5 Hyperparameter.....	59
3.7.6 Determination of Classification Threshold Value	61
3.7.7 Schema Display.....	62
3.8 Browser Extensions System.....	63
3.8.1 System Design.....	64
3.8.2 System Testing.....	66
CHAPTER IV RESULT AND DISCUSSION.....	68
4.1 Dataset Collection	68
4.2 Data Preprocessing.....	69
4.2.1 Data Cleaning.....	69
4.2.2 Feature Extraction	70
4.2.3 Data Splitting	74
4.2.4 Data Normalization	75
4.3 Model Design.....	76
4.3.1 Autoencoder Model.....	76
4.3.2 CatBoost Model	78
4.3.3 Gradient Boosting Model.....	78
4.4 Model Training and Evaluation	79
4.4.1 Autoencoder Model Training.....	80
4.4.2 CatBoost Model Training.....	80
4.4.3 Gradient Boosting Model Training	82
4.4.4 Result Evaluation	84

4.5 Testing and Results Analysis	85
4.5.1 Testing and Result of the Baseline Model	86
4.5.2 Testing and Result of the Autoencoder Model	90
4.5.3 Testing and Result of the Hybrid AE-CatBoost Model	92
4.5.4 Testing and Result of the Hybrid AE-XGBoost Model.....	96
4.5.5 Testing and Result of the Hybrid AE-LightGBM Model	99
4.6 Best Model Configuration.....	103
4.6.1 Best Configuration Evaluation Criteria.....	103
4.6.2 Model Performance Comparison Analysis	103
4.6.3 Selection of the Best Model Configuration.....	105
4.6.4 Details of the Selected Model Configuration.....	106
4.6.5 Determination of Classification Threshold Value	108
4.6.6 Limitations and Practical Implications.....	109
4.7 System Implementation.....	110
4.7.1 System Design Results.....	110
4.7.2 Client-Side Implementation	111
4.7.3 Server-Side Implementation.....	114
4.7.4 System Testing.....	119
4.7.5 False Positive and False Negative.....	131
CHAPTER V CONCLUSION	137
5.1 Conclusion	137
5.2 Recommendations	138
REFERENCES.....	139

LIST OF FIGURES

Figure 2.1 Supervised Learning Algorithm	15
Figure 2.2 Unsupervised Learning Algorithm	16
Figure 2.3 Semi-supervised Learning Algorithm.....	17
Figure 2.4 Autoencoder Structure	22
Figure 2.5 CatBoost Algorithm Structure	29
Figure 3.1 Research Stages	33
Figure 3.2 Data Preprocessing Flow	36
Figure 3.3 Browser Extension System Design.....	64
Figure 4.1 Mendeley Data Dataset.....	68
Figure 4.2 Additional Dataset	69
Figure 4.3 HTML File Dataset.....	70
Figure 4.4 Data Class Distribution Pie Chart.....	70
Figure 4.5 Feature Extraction Result	74
Figure 4.6 Class Distribution for Each Data Proportion.....	74
Figure 4.7 Data After Normalization	76
Figure 4.8 MSE Autoencoder Curve.....	80
Figure 4.9 CatBoost Logloss Curve on 49 Features	81
Figure 4.10 CatBoost Logloss Curve on Averaged Latent Feature Variations	81
Figure 4.11 XGBoost Logloss Curve with 49 Features	82
Figure 4.12 XGBoost Logloss Curve on Average Latent Variation.....	82
Figure 4.13 LightGBM Logloss Curve with 49 Features	83
Figure 4.14 LightGBM Logloss on Latent Variation Averages	83
Figure 4.15 Best Baseline Model Confusion Matrix	87
Figure 4.16 XGBoost Baseline Model (80:10:10 Split) Confusion Matrix	88
Figure 4.17 LightGBM Baseline Model (80:10:10 Split) Confusion Matrix	89
Figure 4.18 Autoencoder Training Epochs on Multiple Latent Dimensions	91
Figure 4.19 Autoencoder MSE Values Trend on Latent Variations.....	92
Figure 4.20 Confusion Matrix Best Autoencoder-Catboost Model.....	94
Figure 4.21 F1-Score of AE-CatBoost Model on Latent Variations	95
Figure 4.22 Best Confusion Matrix Model Autoencoder-XGBoost	98
Figure 4.23 F1-Score of AE-XGBoost Model on Latent Variations	99

Figure 4.24 Confusion Matrix for the Best Autoencoder-LightGBM Model.....	101
Figure 4.25 F1-Score of AE-LightGBM Model on Latent Variations.....	102
Figure 4.26 AE-CatBoost Hybrid Model ROC Curve.....	108
Figure 4.27 Client-Side Folder Structure	110
Figure 4.28 Server-Side Folder Structure	111
Figure 4.29 UI Popup Display During Page Analysis Status.....	113
Figure 4.30 UI Popup Display When a Safe Site Is Detected.....	114
Figure 4.31 UI Popup Display Upon Detection of a Phishing Site.....	114
Figure 4.32 API Headers Display with 200 OK Status.....	122
Figure 4.33 API Payload Display.....	122
Figure 4.34 API Response Display	122
Figure 4. 35 Popup UI Display When Accessing a Phishing Site	123
Figure 4.36 Popup UI Display When Accessing a Legitimate Site	124
Figure 4.37 Site 1 Latency	125
Figure 4.38 Site 2 Latency	125
Figure 4.39 Site 3 Latency	126
Figure 4.40 Site 4 Latency	126
Figure 4.41 Site 5 Latency	127
Figure 4.42 Site 6 Latency	127
Figure 4.43 Site 7 Latency	128
Figure 4.44 Site 8 Latency	128
Figure 4.45 Site 9 Latency	129
Figure 4.46 Site 10 Latency	129
Figure 4.47 Popup UI Display When the Server Is Shut Down.....	130
Figure 4.48 False Positive 1.1	132
Figure 4.49 False Positive 1.2	132
Figure 4.50 False Positive 2.1	133
Figure 4.51 False Positive 2.2	133
Figure 4.52 False Negative 1.1	134
Figure 4.53 False Negative 1.2	135
Figure 4.54 False Negative 2	135
Figure 4.55 False Negative 3	136

LIST OF TABLES

Table 2.1 Previous Studies	5
Table 2.2 Confusion Matrix	18
Table 3.1 URL Dataset Example	34
Table 3.2 HTML Dataset Example	35
Table 3.3 URL Feature Extraction	38
Table 3.4 HTML Feature Extraction.....	40
Table 3.5 URL Feature Extraction Example.....	44
Table 3.6 HTML Feature Extraction Example	45
Table 3.7 Data Normalization Example.....	46
Table 3.8 Data Normalization Result Example.....	47
Table 3.9 Model Evaluation Matrix	56
Table 3. 10 Architecture Comparison	58
Table 3.11 CatBoost Hyperparameter.....	59
Table 3.12 Autoencoder Hyperparameter	60
Table 3.13 Testing Scenarios	62
Table 3.14 Browser Extension System Testing Scenarios.....	66
Table 4.1 Baseline Model Test Results.....	86
Table 4.2 MSE values of Autoencoder on various latent dimensions	91
Table 4.3 Autoencoder-CatBoost Testing Result.....	93
Table 4.4 Autoencoder-XGBoost Testing Result	96
Table 4.5 Autoencoder – LightGBM Testing Result.....	100
Table 4.6 Summary of Best Performance Across All Models	104
Table 4.7 Selected Model Configuration Specifications.....	106
Table 4.8 Black-Box Testing Results for the Browser Extension System.....	120
Table 4.9 Average Latency	130

LIST OF PSEUDOCODES

Pseudocode 1. URL Feature Extraction	72
Pseudocode 2. HTML Feature Extraction.....	73
Pseudocode 3. Data Normalization	75
Pseudocode 4. Building Autoencoder Model	77
Pseudocode 5. Building CatBoost Model	78
Pseudocode 6. Building Gradient Boosting Models	79
Pseudocode 7. Result Evaluation	84
Pseudocode 8. Content Script	112
Pseudocode 9. Background Script	113
Pseudocode 10. Feature Extraction Server Side	116
Pseudocode 11. Data Preprocessing Server Side	117
Pseudocode 12. Model Inference	118
Pseudocode 13. Response Handler	119