



Automated Phishing Detection: A Noval Machine Learning Approach

Ravindra Jayasinghe
Reg. No.: MS21926808

A THESIS
SUBMITTED TO
SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE IN INFORMATION TECHNOLOGY (CYBER SECURITY)

December 2024

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Anuradha Jayakody

Approved for MSc. Research Project:

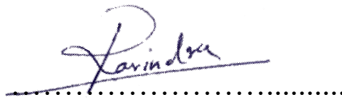
MSc. Programme Co-ordinator, SLIIT

Approved for MSc:

Head of Graduate Studies, FoC, SLIIT

DECLARATION

This is to certify that the work is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources). The work has not previously been submitted in any form to the Sri Lanka Institute of Information Technology or to any other institution for assessment for any other purpose.

A handwritten signature in blue ink, appearing to read 'Ravindra', is written over a horizontal dotted line.

Ravindra Jayasinghe

Date: 12th November 2024.

ABSTRACT

Automated Phishing Detection: A Novel Machine Learning Approach.

Ravindra Jayasinghe

MSc. in Information Technology (Cyber Security)

Supervisor: Prof. Anuradha Jayakody

December 2024

This research contributes a novel machine learning-based approach to cybersecurity, enhancing defenses against phishing and protecting users from emerging online threats. Phishing is an increasingly pervasive cybersecurity threat that exploits user trust by creating fraudulent websites that imitate legitimate ones to steal sensitive information, such as usernames, passwords, and financial details. These deceptive sites use visual and linguistic elements from authentic brands, making them difficult to distinguish from trusted sources and increasing the likelihood of successful attacks. As phishing tactics evolve alongside technological advancements, there is a critical need for robust, adaptive anti-phishing solutions.

This research investigates the application of machine learning to enhance phishing detection, focusing on a model that uses the Gradient Boosting Classifier to identify phishing websites based on key URL features. This approach involves extracting unique characteristics that differentiate phishing URLs from genuine ones, enabling real-time classification and improved detection accuracy. The proposed method systematically analyzes URL features, comparing and contrasting aspects such as domain structure, syntax, and use of brand elements to accurately identify malicious sites.

The model achieved 97.6% accuracy, demonstrating high classification correctness. With a precision of 96.5%, it effectively minimizes false positives, reducing legitimate URL misclassifications. A recall of 98.1% highlights its sensitivity in identifying phishing URLs, and an F1 score of 97.3% balances precision and recall, underscoring its reliability. These results validate the Gradient Boosting Classifier as an effective, adaptable tool against advanced phishing tactics.

ACKNOWLEDGEMENT

While the development of this research proposal was an individual academic endeavor, it would not have been possible without the support and encouragement of many remarkable people, both in my professional and personal life.

First and foremost, I wish to express my heartfelt gratitude to my supervisor, Prof. Anuradha Jayakody, whose guidance and insights were instrumental in helping me achieve my academic goals. His encouragement and expertise were invaluable throughout this journey. I am also deeply appreciative of all the lecturers in my master's program, who not only imparted their knowledge but also shaped my competencies through their dedicated teaching.

I would also like to extend my sincere thanks to the non-academic staff at the Sri Lanka Institute of Information Technology for their assistance with the administrative aspects of the course. Their support ensured that I could focus fully on my academic pursuits.

On a personal note, I am forever grateful to my mother for her unwavering encouragement and belief in me. Her support has been a constant source of strength. I am also profoundly thankful to my wife, who, with tremendous patience and love, took on the responsibility of caring for our baby, allowing me to dedicate the necessary time and focus to this work. Without all their sacrifices and understanding, this accomplishment would not have been possible.

To all of you, I extend my deepest appreciation and heartfelt thanks.

TABLE OF CONTENTS

DECLARATION	ii
ABSTRACT.....	iii
ACKNOWLEDGEMENT	iv
TABLE OF CONTENTS.....	v
List of Figures	viii
List of Tables	ix
Chapter 1 Introduction	1
1.1 Introduction.....	1
1.2 Problem Statement.....	3
1.3 Research Objectives.....	4
1.4 Research Questions.....	5
1.5 Research Significance.....	7
1.6 Research Gap	9
1.7 Scope of the Study	11
1.8 Thesis Organization	11
1.9 Proposed System.....	12
1.9.1 Advantages of Proposed System.....	13
Chapter 2 Literature Review	14
2.1 Organizing the Literature Review.....	14
2.2 Overview of Phishing	15
2.3 Traditional Phishing Detection Techniques.....	19
2.3.1 Blacklisting	19
2.3.2 Heuristic-Based Detection	20
2.3.3 Challenges and Limitations of Traditional Methods.....	21
2.4 Overview of Machine Learning	22
2.4.1 Machine Learning vs. Traditional Programming.....	23
2.4.2 How Machine Learning Works.....	24
2.5 Transition to Machine Learning-Based Detection	26
2.6 Machine Learning in Phishing Detection.....	27
2.6.1 Supervised and Unsupervised Learning Approaches.....	27
2.6.2 Advantages of Machine Learning in Phishing Detection.....	28
2.6.3 Machine Learning Algorithms for Phishing Detection.....	29
2.6.4 Challenges in Applying Machine Learning to Phishing Detection.....	30
2.7 Features and Techniques for Machine Learning-Based Detection	31
2.7.1 URL-Based Attributes.....	32

2.7.2 Domain-Based Attributes.....	32
2.7.3 Page Content-Based Attributes	33
2.7.4 Feature Selection Techniques	34
2.7.5 Impact of Feature Engineering on Model Performances	35
2.8 Algorithmic Approaches in Phishing Detection	36
2.8.1 Logistic Regression.....	36
2.8.2 Decision Trees and Random Forests.....	37
2.8.3 Naïve Bayes	38
2.8.4 Support Vector Machines (SVM)	39
2.8.5 Neural Networks	40
2.8.6 Ensemble Methods: Gradient Boosting and XGBoost.....	40
2.8.7 K-Nearest Neighbors (KNN)	41
2.8.8 Hybrid Approaches and Ensemble Stacking.....	41
2.9 Comparative Studies on Phishing Detection Models.....	42
2.10 Hybrid Frameworks for Enhanced Phishing Detection	43
2.11 Research Gaps in Current Literature.....	45
2.12 Need for a Hybrid Framework in Phishing Detection	47
2.13 Existing Systems Analysis.....	48
2.13.1 Disadvantages of Existing System.....	49
Chapter 3 Methodology	50
3.1 Data Collection	51
3.1.1 Data Collection Process	51
3.1.2 Data Collection Techniques	51
3.1.3 Data Structure	52
3.1.4 Preliminary Data Analysis and Quality Assessment.....	58
3.2 Data Preparation.....	59
3.3 Feature Selection.....	59
3.4 Model Selection	61
3.5 Training and Evaluation.....	62
3.5.1 Cross-Validation	62
3.5.2 Hyperparameter Tuning	62
3.5.3 Evaluation Metrics	63
3.6 Analyze and Prediction	63
3.7 Model Deployment	64
3.8 System Architecture, and Workflow.....	64
Chapter 4 Implementation and Testing	67
4.1 System Requirements.....	67

4.1.1 Hardware Requirements.....	67
4.1.2 Software Requirements	67
4.2 Implementation Steps.....	68
4.2.1 Importing Libraries	68
4.2.2 Loading Data.....	69
4.2.3 Familiarizing with Data & Exploratory Data Analysis (EDA)	69
4.2.4 Visualizing the Data.....	71
4.2.5 Splitting the Data	74
4.2.6 Model Building & Training	75
4.2.7 Gradient Boosting Classifier	76
Chapter 5 Results and Analysis	78
5.1 Performance Analysis	78
5.2 Comparative Study.....	79
5.3 Error Analysis	81
5.4 Implementation Results	82
5.5 Discussion	85
Chapter 6 Conclusion and Future Work	86
6.1 Conclusion	86
6.2 Limitations	86
6.3 Future Work.....	87
References.....	89

List of Figures

Figure 1.1 An illustration of a phishing attack [2].	1
Figure 2.1 Increasing phishing activity since 2020 [19].	17
Figure 2.2 Phishing attacks industry wise[22].	18
Figure 2.3 Traditional Programming	23
Figure 2.4 Machine Learning [36].	24
Figure 2.5 Machine Learning Phase [39].	25
Figure 2.6 Inference from Model [39].	25
Figure 2.7 Logistic regression [48].	36
Figure 2.8 Basic decision tree structure [49].	37
Figure 2.9 Basic Random Forest structure [50].	38
Figure 2.10 3.7.3 Naïve Bayes Classifier [51].	39
Figure 3.1 Parts of a URL [45].	58
Figure 3.2 Proposed system architecture	64
Figure 3.3 Proposed system sequence diagram	65
Figure 3.4 Proposed system data flow diagram	66
Figure 3.5 Proposed system data flow diagram	66
Figure 4.1 Importing required libraries	68
Figure 4.2 Loading data into dataframe.	69
Figure 4.3 Results of Label Classifier.	69
Figure 4.4 Listing the features of the dataset	70
Figure 4.5 Listing information about the dataset	70
Figure 4.6 Visualizing the data 1	71
Figure 4.7 Correlation heatmap	72
Figure 4.8 Visualizing the data 2	72
Figure 4.9 pairplot of the data	73
Figure 4.100 Visualizing the data 3	74
Figure 4.110 Visualizing the data via a pie chart.	74
Figure 4.12 Splitting the dataset in to train and test sets	75
Figure 4.13 Model building and training	76
Figure 4.14 Gradient Boosting Classifier Model	77
Figure 4.15 Predicting the target value from the model for the samples	77
Figure 4.16 Computing the accuracy, f1_score, Recall, precision of the model performance	77
Figure 5.1 Performance Analysis.	78
Figure 5.2 Performance Analysis Results	83
Figure 5.3 URL Prediction Window	83
Figure 5.4 Dataset Upload Window.	84
Figure 5.5 Data Chart Window	85

List of Tables

Table 1.1 Research Gap Comparison [10], [11], [12].....	10
Table 3.1 URL feature set.....	55
Table 5.1 Comparative Study	79