

**AN ACTIVE SELF-LEARNING MODEL FOR
DECEPTIVE PHISHING DETECTION**

Subhash Niroshan Ariyadasa

188077D

Doctor of Philosophy

Department of Computational Mathematics
Faculty of Information Technology

University of Moratuwa
Sri Lanka

July 2023

AN ACTIVE SELF-LEARNING MODEL FOR DECEPTIVE PHISHING DETECTION

Subhash Niroshan Ariyadasa

188077D

Dissertation submitted in partial fulfillment of the requirements for the
degree
Doctor of Philosophy

Department of Computational Mathematics
Faculty of Information Technology

University of Moratuwa
Sri Lanka

July 2023

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: _____

Date: 20-07-2023

The above candidate has carried out research for the PhD Dissertation under my supervision.

Name of the supervisor: Dr. K. S. D. Fernando

Signature of the supervisor: _____

Date: 21/07/2023

Name of the supervisor: Prof. M. S. D. Fernando

Signature of the supervisor: _____

Date: 21-07-2023

DEDICATION

This work is entirely dedicated to my respectful parents and my family. Without their constant support, this dissertation would not have been possible. They always inspire me. At the same time, my special thanks also go to my supervisors, who enlightened me with immense academic knowledge and gave me valuable advice whenever I needed it the most.

ACKNOWLEDGMENTS

In a very special way, I would like to thank the following people who have helped me undertake this research:

- My supervisors, Dr. Subha Fernando and Dr. Shantha Fernando, for their enthusiasm for this research, support, encouragement and patience.
- My progress review examiners, Dr. Chamath Keppitiyagama and Dr. Kasun De Zoysa, for their support, encouragement and feedback.
- My progress review evaluation panel, Dr. Thushari Silva, Dr. Thanuja Sandanayake, Dr. Thilini Piyatilake, Dr. C.R.J. Amalraj and Dr. Sagara Sumathipala, for their support, encouragement and feedback.
- Research coordinators at the Department of Computational Mathematics, Dr. Thushari Silva and Dr. Thilini Piyatilake, for their support when arranging my progress evaluations and other PhD program-related tasks.
- Masters program coordinator, Dr. Sagara Sumathipala, for the support received for my course-work completion.
- Dean, Assistant Registrar, and all other officials at the Faculty of Graduate Studies, the University of Moratuwa for all the support I received during my PhD program.
- Director and other officials at the Center for Information Technology Services (CITeS), University of Moratuwa, for their support in my PhD results evaluation process.
- Vice-Chancellor of Uva Wellassa University, Prof. Jayantha Ratnasekera, for allowing me full-time leaves to complete this PhD program.

- Head of the Department of Computer Science and Informatics, Uva Wellassa University, Ms. Harshani Wickramarathna, for her support and encouragement.
- All the relevant officials at the University of Moratuwa and Uva Wellassa University, for their support.
- My respectful parents, my beloved wife and my loving children, for their constant support and encouragement.
- My brothers, respectful parents-in-law, brother-in-law and sisters-in-law, for their support and encouragement.
- All of my colleagues and friends who helped me throughout this PhD program.

I am incredibly grateful for the immeasurable help and support every one of you has contributed to my program. Words alone cannot express my gratitude to you all. Thank you.

ABSTRACT

Phishing presents an ongoing and dynamic threat to Internet users, targeting personal and confidential information. Existing anti-phishing solutions encounter challenges in keeping up with the ever-changing nature of these attacks, leading to performance degradation over time. This study aims to develop an autonomous anti-phishing solution that effectively counters evolving phishing threats through continuous knowledge updates. To address the challenge of detecting the latest phishing attacks, SmartiPhish, an autonomous anti-phishing solution with continuous learning support, is proposed. Utilizing a quantitative research approach, data is collected from trusted third parties at multiple time points to create a valid dataset. The primary outcome is a reinforcement learning solution that leverages a novel deep learning model alongside Alexa rank and community decisions. The innovative use of Graph Neural Networks in the anti-phishing domain, combined with Long-term Recurrent Convolutional Networks, enables SmartiPhish to estimate a website's phishing probability using URL and HTML content features. Additionally, the study addresses a crucial research gap by developing a reliable method named PhishRepo for collecting and precisely labelling the latest phishing data. SmartiPhish exhibits positive results, achieving a detection accuracy of 96.40%, an f1-score of 96.42%, and an exceptionally low False Negative Rate (FNR) of 0.029. In real-world web environments, the solution outperforms similar solutions and demonstrates enhanced effectiveness against zero-day phishing attacks. Notably, the integration of continuous learning support facilitates a significant 6% improvement in detection accuracy after six weeks. SmartiPhish's adaptive approach integrates a systematic knowledge acquisition process, enabling dynamic updates of phishing detection features to counter the ever-evolving landscape of phishing attacks. The findings highlight its potential in strengthening cybersecurity measures and provide practical insights for dealing with phishing threats in today's digital world. Continuously updating its knowledge base, SmartiPhish stands as a strong defence, promising improved protection for Internet users.

Keywords: Cyberattack, Deep learning, Graph neural networks, Internet security, Reinforcement learning

TABLE OF CONTENTS

Declaration	i
Dedication	ii
Acknowledgments	iii
Abstarct	v
Table of Contents	vi
List of Figures	x
List of Tables	xiii
List of Abbreviations	xiv
1 Introduction	1
1.1 Prolegomena	1
1.2 Background to the study	2
1.3 Research problem	6
1.4 Research aim, objectives and questions	7
1.5 Scope of the study	8
1.6 Significance of the study	8
1.7 Limitations of the study	9
1.8 Structural outline of the dissertation	10
2 Overview of Phishing Attacks	12
2.1 Introduction	12
2.2 Phishing definition	12
2.3 History of phishing	13
2.4 Phishing motives	14
2.5 Phishing medium	15
2.6 Phishing process	15
2.7 Phishing types and techniques	17
2.7.1 Deceptive phishing	17
2.7.2 Technical subterfuge	18
2.8 Mitigation of phishing attacks	20
2.9 Current state of phishing	21

2.10	Summary	23
3	Related Literature	24
3.1	Introduction	24
3.2	Phishing detection	24
3.3	Data collection and labelling	24
3.4	Feature selection and engineering	28
3.5	Detection techniques	32
3.5.1	User education	32
3.5.2	Software-based solutions	35
3.6	Performance evaluation	50
3.7	Present challenges	52
3.8	Problem definition	55
3.9	Summary	56
4	Research Methodology	57
4.1	Introduction	57
4.2	Research design	57
4.3	Solution implementation	61
4.4	Environment setup	67
4.5	Methodological limitations	69
4.6	Summary	70
5	Phishing Detection with URL and HTML	71
5.1	Introduction	71
5.2	Overview of the solution	71
5.2.1	URLDet	72
5.2.2	HTMLDet	76
5.2.3	Deep learning model (DLM)	85
5.2.4	Hybrid DLM	86
5.3	Data collection and preprocessing	91
5.3.1	Classic dataset	93
5.3.2	Modern dataset	95
5.3.3	Benchmark dataset	96
5.3.4	Diversity of datasets	97
5.4	Model training	99
5.4.1	Hybrid DLM training	100
5.4.2	DLM training	100
5.5	Model evaluation	103
5.5.1	Hybrid DLM performance	103
5.5.2	DLM performance	104
5.6	Results and discussion	107
5.7	Summary	110
6	Reinforcement Learning to Enhance Phishing Attack Detection	111
6.1	Introduction	111

6.2	Reinforcement learning (RL)	111
6.3	Reinforcement learning model (RLM)	115
6.3.1	Environment	115
6.3.2	Policy	118
6.3.3	DQN	120
6.3.4	Agent	122
6.3.5	Reward function	123
6.3.6	Phishing detection framework	126
6.4	Phishing detection solution (RDLM)	126
6.5	Data collection and preprocessing	127
6.6	Model training	129
6.7	Model evaluation	130
6.7.1	Overall performance	130
6.7.2	Benchmarking	130
6.8	Results and discussion	131
6.9	Summary	133
7	Phishing Data Collection Process	134
7.1	Introduction	134
7.2	Overview of the proposed process	134
7.2.1	Phishing data collection	136
7.2.2	Labelling process	143
7.2.3	Data dissemination	150
7.3	Target attack prevention (TAP)	153
7.4	Diversity of the collected phishing data	155
7.4.1	Domains distribution and TLDs	157
7.4.2	Distribution of URL character length and HTTPS	158
7.4.3	The tendency of data leakage	159
7.5	Effectiveness of the collected data in machine learning	161
7.5.1	Constructing the datasets	162
7.5.2	Training the solutions	164
7.5.3	Performance evaluation	166
7.6	Discussion	167
7.7	Summary	172
8	Proposed Anti-Phishing Solution	173
8.1	Introduction	173
8.2	Knowledge acquisition process	173
8.2.1	Data production	173
8.2.2	Data submission	174
8.2.3	Labelling	176
8.2.4	Data construction	176
8.2.5	Automatic knowledge acquisition	178
8.3	Autonomous anti-phishing solution	182
8.3.1	SmartiPhish	182
8.3.2	Defense against adversarial attacks	183

8.3.3	Real-time phishing detection	188
8.4	Summary	193
9	Experiments and results	195
9.1	Introduction	195
9.2	Experiments	195
9.2.1	Overall performance	196
9.2.2	Continuous learning ability	197
9.2.3	Zero-day protection	199
9.2.4	Benchmarking	199
9.2.5	Detection time	200
9.2.6	Imbalanced test	201
9.2.7	Real-time phishing detection	202
9.3	Results analysis	204
9.4	Summary	207
10	Evaluation	208
10.1	Introduction	208
10.2	Research overview	208
10.3	Achieving the aim and objectives of the study	209
10.4	Resolving the research problem	226
10.5	Research novelty and contributions	227
10.5.1	Research novelty	227
10.5.2	Main contribution	228
10.5.3	Value-added contributions	229
10.6	Research limitations	232
10.7	Summary	234
11	Conclusion and Recommendations	235
11.1	Introduction	235
11.2	Overall findings	235
11.3	Dissemination of the knowledge	237
11.4	Recommendations for future research	238
11.5	Summary	240
	References	241
	Appendix A. Sample Source Codes	257
	Appendix B. Additional Experiments on Model Selection	293
	Appendix C. Supplementary Information	297

LIST OF FIGURES

1.1	Examples of recent phishing attacks	2
1.2	Information flow of the typical phishing attack	4
2.1	An example of how the correction technique works in phishing mitigation	21
2.2	Number of phishing attacks reported during the last three years	22
3.1	Phishing attack mitigation in terms of phishing detection	25
3.2	Categorisation of phishing detection solutions	32
3.3	Classification confusion matrix	51
4.1	Research methodology	59
4.2	The active learning cycle	66
5.1	Workflow of the URLEDet model	72
5.2	The typical structure of a URL	73
5.3	The URLEDet architecture	74
5.4	Character length distribution of the URLs	74
5.5	The URLEDet model summary	77
5.6	Example of an HTML page in a tree view	79
5.7	Workflow of the HTMLDet model	79
5.8	Example graphs constructed from the graph construction process	82
5.9	The HTMLDet architecture	84
5.10	The HTMLDet model summary	85
5.11	The DLM architecture	86
5.12	A plot of DLM model graph	87
5.13	The URLEDet architecture of the Hybrid DLM	88
5.14	The HTMLDet architecture of the Hybrid DLM	91
5.15	A plot of Hybrid DLM model graph	92
5.16	Distributions of domains and TLDs in the modern and benchmark datasets	98
5.17	Distributions of URL length and HTTPS in the modern and benchmark datasets	99
5.18	Hybrid DLM performance curves	100
5.19	URLEDet performance curves	101
5.20	HTMLDet performance curves	102
5.21	DLM performance curves in phase one training	102
5.22	DLM performance curves in phase two training	103
5.23	DLM detection time curve	107
6.1	The agent–environment interaction in RL	113

6.2	Different types of RL architectures	114
6.3	The proposed RLM architecture	115
6.4	DQN architecture	120
6.5	DQN prediction model	121
6.6	Generated rewards by RLM in different scenarios	125
6.7	Overview of the proposed phishing detection solution	128
6.8	Accumulated mean reward in each episode	130
6.9	Summary of RLM’s prediction results	131
7.1	The landing page of PhishRepo	136
7.2	Workflow diagram of the data collection process	137
7.3	Manual submission interface	139
7.4	Example of a scenario of different URLs for the same phishing target	141
7.5	Estimation of distance threshold d for near-duplicate detection	143
7.6	PhishRepo’s submission labelling process	144
7.7	PhishRepo’s voting interface	147
7.8	Editor’s voting board	149
7.9	Basic details interface of a submission	149
7.10	Commenting pop-up window for phishing votes	149
7.11	Commenting pop-up window for legitimate votes	150
7.12	The hierarchical structure of the zip file	151
7.13	User query interface	152
7.14	Distributions of domains and TLDs in the selected datasets	157
7.15	Phishing URL character length and HTTPS distributions	158
7.16	Percentage Distribution of duplicates and near-duplicates	161
7.17	Example of a near-duplicate found in the PhishRepo dataset	162
7.18	Distributions of legitimate URL character length in selected datasets	164
7.19	Distribution of Accuracy, f1-score, and FNR for the selected solutions under different datasets	168
8.1	Proposed knowledge acquisition process	175
8.2	DLM’s performance during the continuous learning process	179
8.3	SmartiPhish solution	184
8.4	SmartiPhish’s information flow diagram	185
8.5	Overview of a GAN network	186
8.6	SmartiPhish’s daily performance against adversarial attacks	187
8.7	Browser processing pipeline	189
8.8	A REST API-based architecture	190
8.9	MORA browser interface	191
8.10	MORA browser’s ‘Stop Access’ interface	192
8.11	MORA browser’s ‘Ask User’ interface	192
8.12	Proposed phishing detection solution	193
9.1	SmartiPhish performance change overtime	198
9.2	Performance trends of benchmark solutions over 3 months	200
9.3	SmartiPhish detection time curve	201

9.4	Performance comparison with different legitimate to phishing ratios	202
9.5	SmartiPhish's daily performance	203
9.6	SmartiPhish's real-world detection time curve	203

LIST OF TABLES

3.1	Commonly used phishing detection features that come under URL, webpage content and third-party	30
3.2	Overview of the standard phishing detection approaches	36
3.3	Phishing detection solutions exist in the literature	42
3.4	Recent advances in machine learning-based phishing detection solutions	48
5.1	Input X values	83
5.2	Details of the used datasets	93
5.3	Hybrid DLM performance evaluation	104
5.4	The Hybrid DLM comparison with different architectures	104
5.5	DLM performance evaluation	104
5.6	The DLM comparison with selected phishing detection models	105
5.7	The DLM performance evaluation with the benchmark dataset	105
5.8	Results of the zero-day attack detection experiment	106
6.1	Details of the dataset used in RLM training and evaluation	127
6.2	RLM and DLM performance with the test dataset	131
7.1	Used phishing datasets' details	156
7.2	Details of the used machine learning-based solutions	162
7.3	Performance of the trained models with the selected datasets	166
7.4	Comparison of Phisherman and PhishRepo	169
8.1	DLM's learning process	181
8.2	RLM's learning process	183
9.1	RLM instances	196
9.2	SmartiPhish overall performance	196
9.3	Performance fluctuation during a new DLM deployment	198
9.4	SmartiPhish's zero-day detection results	199
9.5	Initial performances of the benchmark solutions	199
9.6	SmartiPhish comparison with selected phishing detection solutions	200

LIST OF ABBREVIATIONS

ACMR	Absolute Cumulative Majority Relabelling
Adam	Adaptive Moment Estimation
AI	Artificial Intelligence
AIWL	Automated Individual White-List
ANN	Artificial Neural Networks
API	Application Programming Interface
ARMA	Auto-Regressive Moving Average
CA	Certificate Authority
CML	Conventional Machine Learning
CNN	Convolutional Neural Network
ComD	Community Decision
CSS	Cascading Style Sheets
Deque	Double-Ended Queue
DL	Deep Learning
DLM	Deep Learning Model
DNS	Domain Name System
DoS	Denial-of-Service
DQN	Deep Q-learning Network
FIFO	First In, First Out
FNR	False Negative Rate
GAN	Generative Adversarial Networks
GBDT	Gradient Boosting Decision Tree
GCN	Graph Convolutional Networks
GCS	Graph Convolutional Skip
GN	Generator Network
GNN	Graph Neural Network
GSB	Google Safe Browsing
HTML	HyperText Markup Language
HTTPS	Hypertext Transfer Protocol Secure
IPR	Initial Phishing Records
KAP	Knowledge Acquisition Process
k-NN	k-Nearest Neighbour
LightGBM	Light Gradient Boosting Machine
LRCN	Long-term Recurrent Convolutional Network
LSTM	Long Short-Term Memory
MDP	Markov Decision Process
MLP	Multi-Layer Perceptron
MPN	Multilayer Perceptron Network

NLP	Natural Language Processing
NN	Neural Networks
PhaaS	Phishing-as-a-Service
pHash	Perceptual Hashing
PNN	Probabilistic Neural Network
ReLU	Rectified Linear Unit
RESTful	Representational State Transfer
RF	Random Forest
RL	Reinforcement Learning
RLM	Reinforcement Learning Model
RoF	Rotation Forest
SaaS	Software as a Service
SEO	Search Engine Optimisation
SGD	Stochastic Gradient Descent
SMO	Sequential Minimal Optimisation
SMS	Short Message Services
SSL	Secure Sockets Layer
SVM	Support Vector Machine
Tanh	Hyperbolic Tangent
TAP	Target Attack Prevention
TF-IDF	Term Frequency-Inverse Document Frequency
TLD	Top-Level Domain
URL	Uniform Resource Locator
XGBoost	eXtreme Gradient Boosting
XSS	Cross-Site Scripting