

DECLARATION

I, Shweta Sankhwar, solemnly declare that this thesis of research on ‘**An Empirical Approach on Detection and Prevention of E-mail Phishing using Machine Learning Techniques**’ is my original work. The study has been conducted under the guidance of Dr. Dharendra Pandey and Prof. R. A. Khan, at Department of Information Technology, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow. It is further declared that to the best of my knowledge and belief it has not been submitted earlier for the award of any degree and also undertaken that the thesis is essentially free from all kinds of plagiarism.

Dated:

(Shweta Sankhwar)
Research Scholar
Department of Information Technology,
Babasaheb Bhimrao Ambedkar University,
(A Central University)
Lucknow-226025, India



बाबासाहेब भीमराव अंबेडकर विश्वविद्यालय

विद्या विहार, रायबरेली रोड, लखनऊ- 226025

BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY
(A Central University)

Vidya Vihar, Rae Bareli Road, Lucknow- 226025

CERTIFICATE

This is to certify that the thesis entitled '**An Empirical Approach on Detection and Prevention of E-mail Phishing using Machine Learning Techniques**' submitted by **Ms. Shweta Sankhwar** is an original research work and has not been previously submitted in part or full for the award of any other degree or diploma to this or any other University. This thesis submitted to Babasaheb Bhimrao Ambedkar University Lucknow satisfies all the requirements as stipulated in the Doctor of Philosophy (Ph.D.) regulations-1999 as amended in 2013 and it is fit for submission and evaluation for the award of the degree of Doctor of Philosophy of the University.

Co-Supervisor

Supervisor

Dated:

Head of the Department

ACKNOWLEDGEMENT

*First and foremost, I would like to express my deep sense of reverence and gratitude, to my supervisor **Dr. Dharendra Pandey** and co-supervisor **Prof. R.A Khan** for his strong support, inspiring guidance and consistent encouragement which has enabled me to achieve my academic goals. His kindness, confidence, and intelligence made all the differences in my academic career. I am blessed to have the supervisors who gave me freedom to open my wings and fly in real-world sky to learn academic as well real-life dimensions without restrictions.*

*I convey my special thanks to **Dr. Arvind Chaturvedi** (IPS. UPP.) inspiring guidance and consistent encouragement. His kind support, motivation and recommendations enhanced me and filled with confidence. I am thankful to **Dr. Shalini Chandra** (Assistant Professor) from core-heart for her insightful comments, encouragement, emotional support and consultations during this endeavor.*

*I am thankful to Director of **ASKY Software.**, allowing me the permission for data collection. I also express my thanks to entire staff member, for supporting me in conducting the experiment. I also acknowledge the help of all **DIT** staff members, for their cooperation and assistance. I am thankful to my all colleagues for their kind support, affection and fun.*

*I am grateful to my parent's **Mrs. Ram Kumari** and **Mr. R.D Sankhwar** who instilled in the passion for knowledge and commitment to set goals and gave me all the emotional and spiritual resources. I would also like to thanks to my guardian **Mr. Satyendra Singh** who provided me with an opportunity and endless support in seeking my dreams. I must also thank to my friends Vinamarta, Sarita, Vanya and Abhimanyu for their unconditional support and encouragement. I special thanks to my friend Dipika and Pooja for their unconditional love, affection and financial support. Above all, I am immensely indebted to God for this achievement.*

Shweta Sankhwar

ABSTRACT

Internet is becoming a significant resource to people in the modern society. To cope-up with the increasing demand for robust and novel web applications, programmers are developing new tricks to add a unique flavour to their web sites. Though these advanced features are based on mature languages and standards, new security problems are often unearthed with each new trick. These new security problems are not only due to technological nuances of new programming techniques, but are also dependent on the way people interact with the web sites. This research work focuses on highlighting one such problem named Phishing, which has become a serious problem for all enterprises utilizing e-communication and online e-commerce tools.

The word phishing is a variation of the word ‘fishing’. The concept of ‘phishing’ can be best understood by observing a typical ‘fishing’ activity, whereby a fisherman offers a tempting bait to the fishes and waits for them to bite onto it hungrily. In the cyber world, ‘phishers’ follow the same methodology; only in this case the bait turns out to be a seemingly harmless email or a text message that demands the receiver’s personal information in lieu of providing him with some ‘too good to be true’ offers. Many people fall prey to such phishing attacks by believing that the messages and emails have originated from some legitimate source and naively reply back with all their sensitive information and credentials. In general terms, these people are said to have been ‘phished’ over the internet.

Phishing websites are cleverly designed replicas of original and trusted websites. Most spoofed websites would seem indistinguishable from their genuine counterparts at first glance. Such websites are dedicated to fraudulent activities and a phishing email is most likely to

contain a link to such a website. When visited, these sites would prompt the user to provide their valuable credentials like passwords or credit card numbers assuring them of huge monetary gains or lucrative offers. Needless to say, this information is utilized by the phishers operating that site for their own benefits. This is precisely why phishing attacks are synonymously termed as ‘identity theft’ attempts.

The problem of phishing attacks in enterprise is next issue rising in wide scale and complexity, as phishers use email phishing via obfuscated, malicious or phished URLs and continuously adapt or innovate their strategies to lure victims. To gain trust and confidence of victim’s phishers have started using visceral factors and Familiarity cues. Although in most cases a phisher’s clear motive is to commit identity theft in order to benefit from it financially; it is wrong to assume that phishing is always money centric. A phisher can also rob an internet user of his goodwill and character. By phishing the login credentials of an employee or a student, a phisher can trespass their accounts and steal their personal information. There are no limits to what a phisher can do in such a scenario. Earning a bad name for oneself in a professional or academic arena can prove much more traumatic than being embarrassed at a social networking site. It is a challenging task to address this issue.

Therefore, in this thesis novel approaches for E-mail Phishing Detection using Machine Learning techniques are proposed. Firstly, A heuristic based model is designed to detect email phishing via URLs (obfuscated, malicious or phished URLs) using Naïve Bayes and Support Vector Machine classifiers. This model includes a URL detection algorithm which efficiently detects phishing and legitimate URLs. It is evident through extensive literature review that single filter approaches are insufficient to detect different categories of phishing attempts in enterprise environ. Therefore, a novel anti-phishing model for enterprise

using artificial neural network is proposed. Basically, it is a heuristic-based approach with ANN. This proposed phishing detection model incorporates an anti-phishing multi-filter for detection phishing/malicious/obfuscated/spoofed URLs. This anti-phishing multi-filter has ability to scan the effective list of URL features with social features or social human factors. In addition, this model effectively identifies whether the phishing email is known phishing or unknown phishing to reduce the trust and familiarity-based email phishing enterprise environ. The Feed-Forward Backpropagation and Levenberg-Marquart methods of ANN are adopted to enhance the URL classification process and with Fuzzy Inference System to get result with imprecise data of social features. The proposed model can effectively handle zero-day phishing and also accurately classify the known and unknown email phishing via URLs. Thirdly, a novel anti-phishing effectiveness evaluator is developed with a formula to calculate the effectiveness of anti-phish armatures or mechanisms. This evaluator calculates the effectiveness based on email structure vulnerability.

Lastly, guidelines for email phishing prevention are chalked out to avoid the phishing attack. The major key point is listed for naive user to recognize the phishing email and to differentiate between phished and legitimate email via URL. Dissemination of cyber awareness and digital hygiene is recommended for the society. Some suggestive measures are also discussed with case studies for prevention and protection of user from cyber-crime. As it would keep them secured in cyber, social and monitory aspects and also would help them spread the cyber awareness through publicity or campaigning.

ABBREVIATIONS

ANN	: Artificial Neural Network
APWG	: Anti- Phishing Working Group
DNS	: Domain Name System
FBNN	: Feed-Forward Backpropagation Neural Network
FIS	: Fuzzy Inference System
FLC	: Fuzzy Logic Control
FPR	: False Positive Rate
FRBS	: Fuzzy Rule Based System
HTTP	: Hyper Text Transfer Protocol Secure
IP Address	: Internet Protocol Address
KB	: Knowledge Base
LM	: Levenberg-Marquardt neural network
MF	: Membership Function
ML	: Machine Learning Techniques
NB	: Naïve Bayes
RMSE	: Root Mean Square Error
SVM	: Support Vector Machine
TLD	: Top Level Domain
TLDs	: Top Level Domains
TPR	: True Positive Rate
URL	: Uniform Resource Locator

TABLE OF CONTENT

Declaration		i
Certificate		ii
Acknowledgment		iii
Abstract		iv
Abbreviations		vii
List of Figures		xiii
List of Tables		xv
S.N.	Topic	Page No.
	CHAPTER 1: PROLOGUE	1-7
1.1	Introduction	1
1.2	Motivation for the Research	2
1.3	Research Problem	4
1.4	Objectives of the Research	5
1.5	Expected Deliverables	6
1.6	Thesis Outline	6
	CHAPTER 2: BACKGROUND INFORMATION AND LITERATURE REVIEW	8-23
2.1	Introduction	8
2.2	Background	8
2.3	Evolution of Phishing	11
2.3.1	Top Most Targeted Industry Sector	12
2.3.2	Top Most Port Number used in Phishing	13
2.3.3	Top Level Domain (TLDs) used to Host Phishing Site	13
2.4	State-of-art of E-mail Phishing	15
2.5	A Case Study of E-mail Phishing	16

2.6	E-mail Phishing in Enterprise Environ	18
2.7	Available Anti-Phish Approaches and E-mail Phishing Classification Research Trends	19
2.8	Conclusion	22
CHAPTER 3: A PHISHING DETECTION MODEL USING SUPERVISED LEARNING ALGORITHMS		24-45
3.1	Introduction	24
3.2	URL Structure	25
3.3	Proposed Approach for E-mail Phishing Detection	26
3.3.1	Architecture of Proposed Model	26
3.3.2	URL Feature Set	27
3.4	Enhanced Malicious URL Detection Algorithm	32
3.5	Implementation of the Proposed Model	32
3.6	Experimental Setup with Dataset-I	34
3.6.1	Performance Analysis	35
3.6.2	Performance Evaluation with Naïve Bayes Classifier	36
3.6.3	Performance Metric and Comparative study	39
3.7	Experimental Setup with Dataset-II	40
3.7.1	Performance Evaluation with Support Vector Machine	41
3.7.2	Performance Metric	42
3.7.3	Comparative Study of Proposed and Existing Model	44
3.8	Conclusion	44
		46-85

**CHAPTER 4: ANTI-PHISHING IN ENTERPRISE
ENVIRON USING ARTIFICIAL NEURAL NETWORK**

4.1	Introduction	46
4.2	Phishing- An Enterprise Threat	47
4.2.1	Phishing Attack through Social Engineering	48
4.2.2	Phishing Strategies in Enterprise Environ	49
4.3	Social Facets Triggering Phishing Attacks in Enterprise Environ	50
4.4	Proposed Approach for E-mail Phishing Detection	51
4.4.1	Architecture of Proposed Approach	53
4.4.2	URL Feature Set	55
4.4.3	Social Feature Set	58
4.4.4	Anti-Phishing Multi-Filter	59
4.5	Basic Principle of Artificial Neural Network	61
4.5.1	Feed-Forward Backpropagation Neural Network	61
4.5.2	Levenberg-Marquardt Neural Network	62
4.6	Implementation of the Proposed Approach Phase-I	63
4.6.1	Dataset	63
4.6.2	Performance Evaluation	63
4.6.3	Result and Discussion	65
4.7	Implementation of the Proposed Approach Phase-II	65
4.7.1	Dataset	66
4.8	Fuzzy Rule-Based System (FRBS)	66
4.9	Fuzzy Linguistic variable and Membership Function	67
4.9.1	Description of Fuzzy input variable	68
4.9.2	Description of Fuzzy output variable	69
4.9.3	Determining Fuzzy Rule Base from input and output variables	70
4.9.4	Fuzzy Rules and Coding	70

4.10	Working Principle of Traditional FLC (Mamdani Approach)	70
4.11	Result and Discussion	72
4.12	Phishing Prevention Guidelines	74
4.12.1	Cyber Awareness and Hygiene	74
4.12.2	Phishing Prevention on Ground Level	75
4.12.3	Phishing Precautionary Measures at Enterprise Environ	77
4.12.4	Sturdy & Meticulous Web Development is Recommended	78
4.12.5	Suggestive Measures for other Cyber-crime	78
4.13	Implementation of Phishing Prevention Guidelines	79
4.14	Validation of Phishing Prevention Guidelines	82
4.15	Conclusion	84

CHAPTER 5: ANTI-PHISHING EFFECTIVENESS EVALUATOR MODEL 86-98

5.1	Introduction	86
5.2	Proposed Approach for Anti-Phish Effectiveness Evaluation	86
5.2.1	Architecture of proposed Approach	87
5.3	Implementation of Proposed Approach	88
5.3.1	Effectiveness Evaluation of Existing and Proposed Approaches	93
5.4	Results and Discussion	95
5.5	Conclusion	97

99-104

CHAPTER 6: EPILOGUE

6.1	Summary of Thesis	99
6.2	Research Contribution	101
6.3	Significance of the Study	103
6.4	Limitation	104
6.5	Agenda for Future Research	104

REFERENCES	105-117
-------------------	----------------

ANNEXURES	118-127
------------------	----------------

LIST OF FIGURES

Figure No.	Figure Name	Page No.
Figure 2.2 (a)	: User's Satisfaction with Internet Banking	10
Figure 2.2 (b)	: Problems occurs in Internet Banking	10
Figure 2.3 (a)	: Top Most Targeted Industry Sectors	12
Figure 2.3 (b)	: Top used Ports Hosting Phishing Data Collection Servers	14
Figure 2.3 (c)	: Top most TLDs used to Host Phishing Site	14
Figure 2.4	: State-of-art of URL based Email Phishing	15
Figure 2.5 (a)	: Phishing Webpage	16
Figure 2.5 (b)	: Phishing URL	17
Figure 3.2	: URL Structure	25
Figure 3.3	: Architecture of Enhanced Malicious URL Detection Model	27
Figure 3.4	: Enhanced Malicious URL Detection (EMUD) Algorithm	33
Figure 3.7	: Support Vector Machine	41
Figure 4.4 (a)	: Architecture of Anti-PhiEE Model	52
Figure 4.4 (b)	: Design of Anti-Phishing Multi Filter (APMF)	54
Figure 4.4 (c)	: Anti-Phishing Multi-Filter (APMF) Algorithm	60
Figure 4.5 (a)	: 25X2 Input Output Neural Network Architecture	62
Figure 4.5 (b)	: Artificial Neural Network	62
Figure 4.6 (a)	: Performance of Feed-Forward Backpropagation Neural Network	64
Figure 4.6 (b)	: Performance of Levenberg-Marquardt Neural Network	64
Figure 4.8 (a)	: A schematic view of an FRBS	67
Figure 4.9 (a)	: Membership Function Distributions for the Variables	68

Figure.4.9 (b) : Membership Function Distributions for Output Fuzzy Variable	69
Figure 4.11 (a) : Social Media Contacts (X_1) vs Social Media Common Contacts (X_2)	73
Figure 4.11 (b) : Social Media Contacts (X_1) vs Social Media Common Activity (X_3)	73
Figure 4.11 (c) : Social Media Common Contacts (X_2) vs Common Activity in Social Media (X_3)	74
Figure 5.2 : Anti-Phishing Effectiveness Evaluator Model (APEE Model)	87
Figure 5.4 (a) : A Slew of Phishing Vulnerabilities	95
Figure 5.4 (b) : Effectiveness of Existing Anti-Phishing Mechanisms	96

LIST OF TABLES

Table No.	Table Name	Page No.
Table 2.1	Available Anti-Phish Approaches	20
Table 3.6	: List of URLs	34
Table 3.6.1 (a)	: Experimental Results of EMUD Algorithm	35
Table 3.6.1 (b)	: Experimental Results of EPCMU Algorithm	35
Table 3.6.2	: Naïve Bayes Classification of URLs	37
Table 3.6.3	: Confusion matrix of EMUD and Existing Model	40
Table 3.7.2	: Performance for Proposed Model	43
Table 3.7.3	: Comparison of Proposed Model with Existing Model	44
Table 4.4.2 (a)	: Suspicious URL Forms or Patterns	55
Table 4.4.2 (b)	: URL Property Values	56
Table 4.4.2 (c)	: Page Ranking	56
Table 4.4.2 (d)	: Google suggestion for URL Authenticity	57
Table 4.6.3	: Comparison between FFNN and LM Neural Network	65
Table 4.9.1	: Linguistic term and their range	69
Table 4.9.4	: Description of fuzzy linguistic term	70
Table 4.13	: Phishing Prevention Guideline Weight	79
Table 4.14	Validation of Phishing Prevention Guidelines	83
Table 5.3	: A Slew of E-mail Phishing Vulnerabilities	90
Table 5.3.1 (a)	: Effectiveness Evaluation Anti-Phishing Mechanisms	93
Table 5.3.1 (b)	: Domain-based Effectiveness Evaluation of Proposed Anti-Phishing Mechanism	94