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. Dhirendra 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

BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY
LUCKNOW
प्रज्ञा शील करुणा
ESTABLISHED 1996

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. Dhirendra 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 s 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

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

CHAPTER 6: EPILOGUE

LIST OF FIGURES

LIST OF TABLES