Implementing Advance Machine Learning Algorithms for Detection of phishing webpages in real world situations

Mohit Tiwari [#1], Aashima Airan [*2], Arjun Sinha [#3], Devansh Rastogi [#4]
Himanshu Dadhwal [#5] , Kanishk Kumar [#6], Pranav Gaur [#7]

[#]Computer Science and Engineering Department

[*]Electrical and Electronics Engineering Department

[1] mohit.tiwari@bharatividyapeeth.edu

[2] ashimajain046@gmail.com

[3] arjunsinha141@gmail.com

[4] devanshrastogi2407@gmail.com

[5] himanshu2002dadhwal@gmail.com

[6] kumark_anishk03@yahoo.in

[7] pranavgaur354@gmail.com

Abstract: With the ever-growing reliance on online platforms for communication, commerce, and information retrieval, the threat of phishing attacks has become a pervasive concern in the realm of cybersecurity. Phishing, typically a cybercrime, is a deceptive practice where attackers mimic trustworthy entities to trick users into divulging sensitive information, poses a significant risk to individuals and organizations alike. This research paper explores the utilization of Machine Learning (ML) techniques for the real-world detection of phishing webpages, aiming to fortify cyber defences against this evolving threat. Certain machine learning techniques, primarily Random Forest, XGBoost, Support Vector Machine and Decision Tree will be implemented and their accuracy will be noted in view of developing an accurate model.

Keywords: Phishing, cybersecurity, cybercrime, Machine Learning, sensitive.

I. INTRODUCTION

In an era dominated by digital connectivity and online interactions, the pervasive threat of phishing attacks has emerged as a formidable challenge to cybersecurity. Phishing, a malicious practice wherein attackers impersonate trusted entities to deceive users into divulging sensitive information, continues to evolve in sophistication and prevalence. As individuals and organizations increasingly rely on digital platforms for communication, financial transactions, and information sharing, the need for robust and adaptive defense mechanisms against phishing attacks becomes paramount.

Traditional methods of phishing detection often struggle to keep pace with the dynamic and cunning tactics employed by cybercriminals. The static nature of rule-based systems and signature-based approaches falls short in addressing the intricacies of modern phishing campaigns. In response to this growing challenge, the integration of advanced technologies, particularly Machine Learning (ML), has emerged as a promising avenue for enhancing the detection capabilities in real-world situations.

This research embarks on a journey to explore the utilization of ML techniques as a proactive and dynamic solution for the detection of phishing webpages. The focus extends beyond theoretical frameworks to practical applications, considering the complex and dynamic nature of the online environment where phishing attacks unfold. The aim is to develop a robust system capable of adapting to the evolving tactics of cyber adversaries, thereby fortifying the resilience of cybersecurity measures in the face of an ever-expanding threat landscape.

The escalating sophistication of phishing attacks necessitates a departure from traditional security paradigms, prompting a shift towards intelligent, learning-based approaches. Machine Learning, with its ability to analyze patterns, extract features, and learn from historical data, offers a promising avenue to tackle the challenges posed by dynamic and stealthy phishing campaigns. This research seeks to bridge the gap between theoretical advancements in ML and their practical implementation for real-world detection, thereby contributing to the development of effective countermeasures against phishing threats.

As we delve into the intricacies of ML applications for phishing detection, the following sections will elaborate on the selection and implementation of ML algorithms, the importance of feature extraction techniques and the significance of diverse and high-quality datasets. The ultimate goal is to present a comprehensive understanding of how ML can be effectively leveraged to detect phishing webpages in the complex and dynamic landscape of real-world cybersecurity.

II. PHISHING TECHNIQUES

In this section we will deal with some of the commonly used phishing techniques.

a) Email Phishing:

Attackers send emails that appear to be from a legitimate source, with the goal of tricking the recipient into providing sensitive information. They may use email addresses that mimic those of trusted entities, making it difficult for users to discern the phishing attempt or they even can show them as trusted individuals to request sensitive information.

b) Spear Phishing:

Phishers customize their messages for specific individuals or organizations, often using information gathered from social media or other sources to make the emails more convincing. They may impersonate executives, managers, or other high-authority figures within an organization to increase the likelihood of success.

c) Vishing (Voice Phishing):

Attackers use phone calls to trick individuals into providing sensitive information or performing actions, such as revealing account credentials or making financial transactions. For this phishers manipulate caller ID information to that the call is coming from a trusted source.

d) Smishing (SMS Phishing):

Attackers send deceptive text messages, often containing links or instructions, to trick individuals into divulging information or clicking on malicious links. When the user click on the link some sort of malware is installed on their system.

e) Man-in-the-Middle (MitM) Attacks:

Attackers intercept communication between the user and a trusted entity, allowing them to eavesdrop on sensitive information. Sometimes Phishers take control of an ongoing session between the user and a legitimate website to capture sensitive data.

f) Clone Phishing:

Attackers create replica websites that closely mimic legitimate ones, tricking users into entering their credentials, which are then stolen.

g) Search Engine Phishing:

Hackers tamper with search results, hiding malicious sites within seemingly normal links using SEO. These sites can steal your information through phishing.

h) Credential Harvesting:

Phishers create fake login pages that closely resemble legitimate login portals to capture user credentials. Malicious software like keylogger is used to record keystrokes, capturing login credentials and other sensitive information. This leaks all the personal information of a user.

III. LITERATURE REVIEW

AUTHOR NAME	TITLE	APPROACH USED	RESEARCH GAP	FOCUS	ACCURACY
Ping Yi , Yuxiang Guan, Futai Zou, Yao Yao, Wei Wang, and Ting Zhu	Web Phishing Detection Using a Deep Learning Framework	Deep Belief Network (DBN) and direct checking of URL, and age of the domain.	We are primarily focused on Random Forest algorithm	Analyzing the features of phishing websites and present two types of feature for web phishing detection. Introduction of DBN to detect phishing websites and discuss the detection model and algorithm for DBN.	DBN: 90%
Meenu , Sunila godara	Phishing Detection using Machine Learning Techniques	Hybrid approach using Logistic Regression (LR), Support Vector Machines (SVM), Decision Tree (DT), and Neural Networks (NNet)	Author primarily focuses on prediction of phishing emails only.	To construct a spam channel utilizing various machine learning techniques for predicting phishing emails and improves logistic regression technique by using feature selection methods and improves the accuracy to detect phishing.	Logistic regression: 94.1% Neural network: 94.31% Decision tree: 93.9% Support vector machine: 90.1% Improved logistic regression: 95.55%
Trevor Wooda, Vitor Basto-Fernandesb , Eerke Boitenc , Iryna Yevseyevad	Anti-Phishing Defences and Their Application to Before-the-click Phishing Email Detection	General analysis of different defences deployed to counter phishing attacks in particular attacks targeting Company and organisations like Spear phishing, whaling.	The authors are concentrated only on the fact that most of the phishing attacks wants the user to click a link.	To do a in depth analysis of which is the best method suited to prevent before-click event, spear phishing, whaling etc.	N/A
Dinesh P., Mukesh, Navaneethan, Sabeenian R., Paramasivam M., Manjunathan	Identification of Phishing Attacks using Machine Learning Algorithm	Random Forest, XGBoost, and Logistic Regression.	The research focuses only on URL based phishing attack, the dataset collected was small and the storage usage was more.	The aim of this project is to use the dataset produced to predict phishing websites to build machine learning algorithms and deep neural networks.	XGBoost Classifier: 94.2%

Vahid Shahrivari, Mohammad Mahdi Darabi, Mohammad Izadi	Phishing Detection Using Machine Learning Techniques	Logistic Regression, KNN,, Decision Tree, Random Forest, XGBoost	Combination of Several Classifiers had produced insufficient results.	Detection of phishing website using ML as they have some common characteristics between them.	LR: 92.65%, KNN: 96.29% Decision Tree: 96.59% Random Forest: 97.26% XGBoost: 98.32%
Tariro Manyumwa, Phillip Francis, Hanlu Wu, Shouling Ji	Towards Fighting Cybercrime: Malicious URL Attack Type Detection Using Multiclass Classification	Detection of malicious URL attack types using XGBoost, LightGBM, CatBoost.	The Research explores the URL feature based detection only.	To address the detection of malicious URLs in a multiclass classification setting. It focuses on three common URL attack types: phishing, spam, and malware.	XGBoost: 93%
Saleem Raja A., Vinodini R., Kavitha A.	Lexical features based malicious URL detection using machine learning techniques	The proposed method in the research consists two phases, feature extraction and feature reduction.	Other classification of malicious URLs such as shortened link is not addressed.	The paper presents the light weighted method which includes only lexical features of the URL which consists of three phases, Feature Extraction, Feature reduction, Training the Model.	KNN gives better results when considering execution time and accuracy and then Random Forest
Neelam Yadav & Supriya P. Panda	Feature Selection for email phishing detection using machine learning	Random Forest	Approach used by the authors seems different with respect to the approach used by us.	Using machine learning algorithm such as random forest for analysis and classification of emails under category of phishing.	Random Forest achieved accuracy of 99%
Matte Revanth Hari Narayana, Dr.G.K. Sandhia	Phishing Detection using Machine Learning	Naïve Bayes	We are primarily focused on Random Forest algorithm	Detection of phishing website using ML as they have some common characteristics between them.	Naïve Bayes: 98.33%
Mohammed Abutaha, Mohammad Ababneh, Khaled A. Mahmoud, Sherenaz W. Al-Haj Baddar	URL Phishing Detection using Machine Learning Techniques	Random Forest, Gradient Boosting, Neural Network and Support Vector Machine	We are not just focused on phishing detection using URL	To analyse and address URLs as malicious or not, focusing on various machine learning algorithm.	Support Vector Machine (SVM): 99.89% (highest)

	based on URLs Lexical Analysis				
Junaid Rashid, Toqeer Mahmood, Muhammad Wasif Nisar, Tahira Nazir	Phishing Detection Using ML	Random Forest and Support Vector Machine	Technique used by author is different from that used by us	The research uses various machine learning models and algorithms in view to classify the websites as potential phishing site or not.	Random Forest: 94.27%, Support Vector Machine: 95.66%
Rana Abdulraheem, Ammar Odeh, Mustafa Al Fayoumi, Ismail Keshta	Efficient Email phishing detection using Machine learning	Logistic Model Tree	We are primarily focused on Random Forest algorithm	Study and Classification of emails under category of phishing using machine learning techniques.	Logistic Model Tree: 96.9245%

IV. METHODOLOGY

1. Data Acquisition and Preprocessing

The research utilises five different types of data sets, which are publicly available on Github. The set of phishing URLs are collected from opensource service called PhishTank. This service provide a set of phishing URLs in multiple formats like csv, json etc. that gets updated hourly. (https://www.phishtank.com/developer_info.php). From this dataset, 5000 random phishing URLs are collected to train the ML models.

The legitimate URLs are obtained from the open datasets of the University of New Brunswick,https://www.unb.ca/cic/datasets/url-2016.html. This dataset has a collection of benign, spam, phishing, malware & defacement URLs. Out of all these types, the benign url dataset is considered for this project. From this dataset, 5000 random legitimate URLs are collected to train the ML models.

2. Model Building and Training

(A) Random Forest: Ensemble method utilizing decision trees, diversifies feature selection to prevent overfitting and bolster classification robustness.

(B) Decision Tree: Hierarchical model mimicking a flowchart. Asks a series of feature-based questions (e.g., URL length) to reach a final classification (phishing or legitimate). Offers clear reasoning behind its decisions.

(C) XGBoost: A powerful machine learning technique that builds on decision trees, adding a "boosting" step to improve accuracy and handle complex phishing detection tasks.

(D) Multilayer Perceptrons: Artificial neural networks with stacked layers of processing units. Used for complex pattern recognition in data, like identifying phishing attempts based on website features.

(E) Autoencoder Neural Networks: Neural networks that learn efficient data representations. Can identify inconsistencies in website content, potentially uncovering phishing attempts by highlighting unusual patterns.

(F) Support Vector Machines: Finds an optimal boundary (hyperplane) in complex data (phishing vs. legitimate) to achieve the best separation, focusing on critical data points for robust classification.

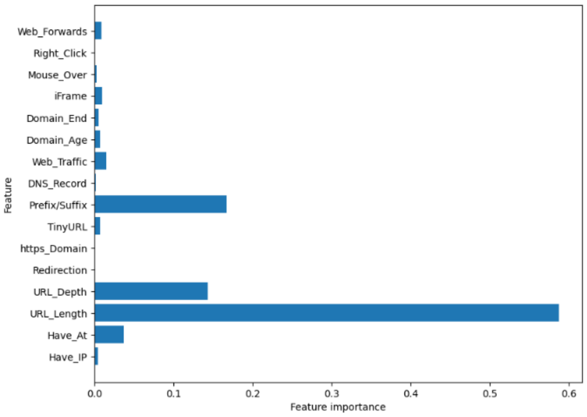


Fig.1. Random Forest

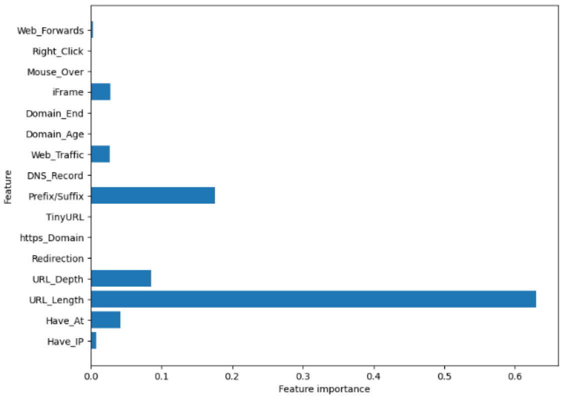
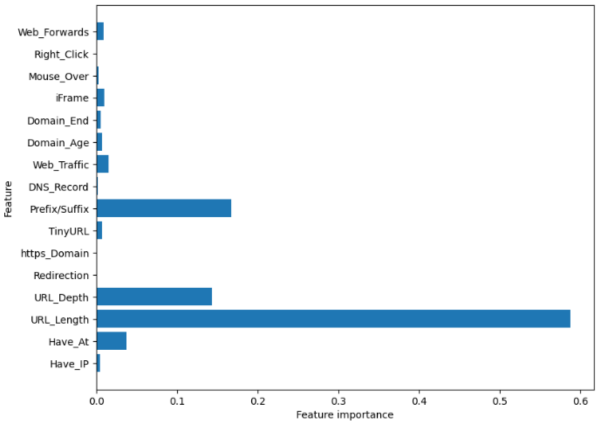


Fig.2. Decision Tree



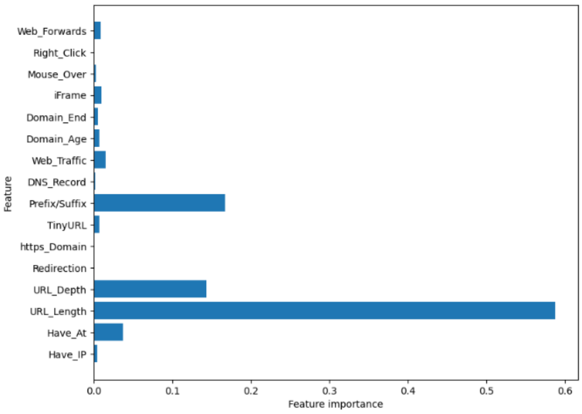
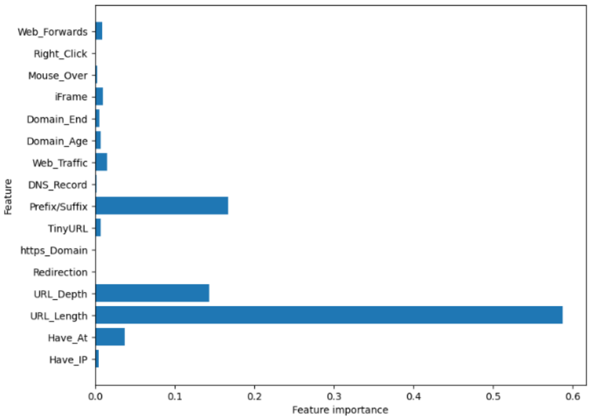


Fig.3. XGBoost

Fig.4. Multilayer Perceptron



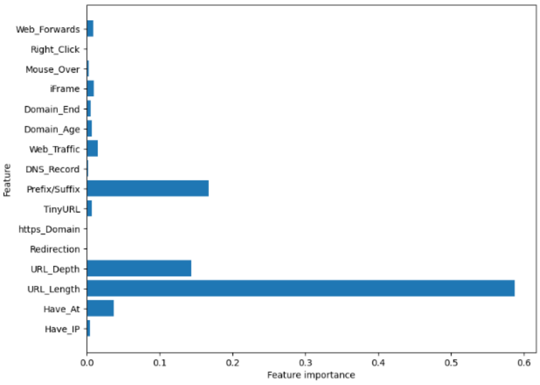
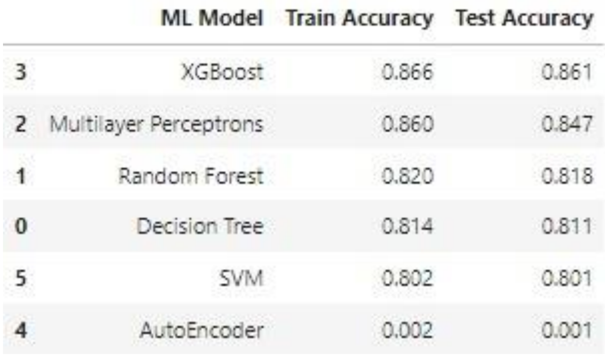


Fig.5. Autoencoder Neural Networks

Fig.6. Support Vector Machine

V. RESULT AND CONCLUSION

This investigation endeavored to leverage cutting-edge machine learning algorithms for the purpose of identifying fraudulent webpages across a spectrum of five distinct datasets. The research employed a battery of six machine learning models, including Random Forests, Decision Trees, Support Vector Machines, Autoencoder Neural Networks, Multilayer Perceptrons, and XGBoost. Within the provided datasets, XGBoost exhibited the most proficient performance, achieving a peak training accuracy of 86.6% and a pinnacle testing accuracy of 86.1%. This was followed closely by Multilayer Perceptrons, which garnered an accuracy of 86% during training and 84.7% during testing.



	ML Model	Train Accuracy	Test Accuracy
3	XGBoost	0.866	0.861
2	Multilayer Perceptrons	0.860	0.847
1	Random Forest	0.820	0.818
0	Decision Tree	0.814	0.811
5	SVM	0.802	0.801
4	AutoEncoder	0.002	0.001

Fig.7. Different Algorithms Result

VI. REFERENCES

[1] A. F. Nugraha, D. A. Tama, D. A. Istiqomah, S. T. A. Ramadhani, B. N. Kusuma and V. A. Windarni. "Feature Selection Technique for improving classification performance in the web-phishing detection process". Jan. 2022.

[2] G. Gressel and K. Achuthan. "Feature Selection for Phishing Detection with Machine Learning". Nov. 2019.

[3] "Tuning the False Positive Rate / False Negative Rate with Phishing Detection Models". Dec. 2019.

[4] Abdulhamit Subasi, Esraa Molah, Fatin Almkallawi, Touseef J. Chaudhery, “Intelligent phishing website detection using Random Forest classifier,” International Conference on Electrical and Computing Technologies and Applications(ICECTA), 2017

[5] Joshi, A., Pattanshetti, P., & Tanuja, R. (2019). Phishing attack detection using feature selection techniques. In International conference on communication and information processing (ICCIP), Nutan College of Engineering and Research.

[6] Ping Yi , Yuxiang Guan, Futai Zou, Yao Yao, Wei Wang, Ting Zhu. “Web Phishing Detection Using a Deep Learning Framework.” 26 Sept. 2018.

[7] Meenu , Sunila godara. "Phishing Detection using Machine Learning Techniques" 30 December. 2019.

[8] El Aassal, A., Baki, S., Das, A., & Verma, R. M. (2020). An indepth benchmarking and evaluation of phishing detection research for security needs. IEEE Access, 8, 22170–22192.

[9] Feng, Q., Tseng, K. K., Pan, J. S., Cheng, P., & Chen, C. (2011). New anti-phishing method with two types of passwords in openid system. In 2011 Fifth international conference on genetic and evolutionary computing (pp. 69–72). IEEE

[10] Hulten, G. J., Rehfuss, P. S., Rounthwaite, R., Goodman, J. T., Seshadrinathan, G., Penta, A. P., Mishra, M., Deyo, R. C., Haber, E. J., & Snelling, D. A. W. et al. (2014). Finding phishing sites. US Patent 8,839,418.

[11] Hutchinson, S., Zhang, Z., & Liu, Q. (2018). Detecting phishing websites with random forest. In International conference on machine learning and intelligent communications (pp. 470–479). Springer.

[12] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A Comprehensive Survey of AI-enabled Phishing Attacks Detection Techniques," Telecommunication Systems, vol. 76, pp. 139–154, 2021, doi: 10.1007/s11235-020-00733-2.

[13] F. Salahdine, Z. El Mrabet, and N. Kaabouch, "Phishing Attacks Detection: A Machine Learning-Based Approach," in Proceedings of the Ubiquitous Computing, Electronics & Mobile Communication Conference, 2021, doi: 10.1109/UEMCON53757.2021.9666627.

[14] P. Yi, Y. Guan, F. Zou, Y. Yao, W. Wang, and T. Zhu, "Web Phishing Detection Using a Deep Learning Framework," Wireless Communications and Mobile Computing, 2018, doi: 10.1155/2018/4678746.

[15] R. Jha and G. Kunwar, "Machine Learning-based URL Analysis for Phishing Detection," presented at the 2023 6th International Conference on Information Systems and Computer Networks (ISCON), 2023, doi: 10.1109/ISCON57294.2023.10112057.

[16] Park G, Stuart LM, Taylor JM, Raskin V. Comparing machine and human ability to detect phishing emails. In: 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 2014. p. 2322–7.

[17] Basto-Fernandes V, Yevseyeva I, Méndez JR, Zhao J, Fdez-Riverola F, T.M. Emmerich M. A spam filtering multi-objective optimization study covering parsimony maximization and three-way classification. Appl Soft Comput. 2016 Nov 1;48:111–23.

[18] Abdulhamit Subasi, Esraa Molah, Fatin Almkallawi, "Intelligent Phishing Website Detection using Random Forest Classifier," ," presented at the 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC).

[19] Ruano-Ordás D, Basto-Fernandes V, Yevseyeva I, Méndez JR. Evolutionary Multi-objective Scheduling for Anti-Spam Filtering Throughput Optimization. In: Martínez de Pisón FJ, Urraca R, Quintián H, Corchado E, editors. Hybrid Artificial Intelligent Systems. Cham: Springer International Publishing; 2017. p. 137–48. (Lecture Notes in Computer Science).

[20] Yevseyeva I, Basto-Fernandes V, Ruano-Ordás D, Méndez JR. Optimising anti-spam filters with evolutionary algorithms. Expert Syst Appl. 2013 Aug 1;40(10):4010–21.

[21] Ala Mughaid,Shadi AlZu'bi, Adnan Hnaif,Salah Taamneh, Asma Alnajjar and Esraa Abu Elsoud."An intelligent cyber security phishing detection system using deep learning techniques". 14 May 2022.

[22] Samuel Marchal and N. Asokan."On Designing and Evaluating Phishing Webpage Detection Techniques for the Real World".2018

[23] Meenu , Sunila godara"An enhanced phishing email detection model using machine learning techniques" international journal of emerging technologies and innovative research 11 ,vol 5,pp523-529 , November 2018.

[24] M. Abutaha, M. Ababneh, K. Mahmoud and S. A. -H. Baddar, "URL Phishing Detection using Machine Learning Techniques based on URLs Lexical Analysis," 2021 12th International Conference on Information and Communication Systems (ICICS), Valencia, Spain, 2021, pp. 147-152, doi: 10.1109/ICICS52457.2021.9464539.

[25] Salloum, "Phishing Email Detection using Natural Language Processing Techniques: A Literature Survey," Procedia Computer Science, 2021.

[26] Kang Leng Chiew, Choon Lin Tan, Kok Sheik Wong, Kelvin S.C. Yong, Wei King Tiong "A New Hybrid Ensemble Feature Selection Framework for Machine Learning-Based Phishing Detection System," Procedia Computer Science , 2021.