Phishing Detection: A Multilayer Approach to Scale Down Phishing

Reddyvari Venkateswara Reddy
Associate Professor, Department of CSE(Cyber Security), CMR College Of Engineering & Technology, Hyderabad, Telangana
Email: venkatreddyvari@cmrcet.ac.in
Punyaban Patel
Professor, Department of CSE(Cyber Security), CMR College Of Engineering & Technology, Hyderabad, Telangana
Email: punyaban@gmail.com
Garikapati Charan Sai
Department of CSE(Cyber Security), CMR College Of Engineering & Technology, Hyderabad, Telangana
Email: charansaigarikapati@gmail.com
Akula Rishika
Department of CSE(Cyber Security), CMR College Of Engineering & Technology, Hyderabad, Telangana
Email: akularishika10@gmail.com
Garlapati Dheeraj Reddy
Department of CSE(Cyber Security), CMR College Of Engineering & Technology, Hyderabad, Telangana
Email: dheerajreddydd@gmail.com

----------ABSTRACT----------

Phishing detection is a process to detect phishing attacks before they happen, inform administrators and users about them, and most importantly reduce the risk. Phishing is a form of cybercrime in which hackers attempt to trick victims into revealing confidential information, e.g. passwords or financial details, by pretending that they're an efficient source for the attack via email and most commonly text messaging. Since phishing is mostly a semantics-based attack that targets human vulnerabilities, identifying these phishing websites can be difficult. The authors suggested the Phishing Detection project as a solution to this problem. By identifying and quickly warning users of potential phishing and malware threats, our project aims to improve online security. This all-inclusive method includes Greek alphabet analysis, port forwarding detection, database comparison, grammar analysis utilizing NLP libraries, and many more. A modified version of phishing detection has been proposed which integrates Greek alphabet analysis , port forwarding detection and homograph attack detection.

Keywords – **Confidential information, Homograph attack, Greek alphabet analysis, Phishing attacks, Port forwarding, Semantics-based attack.**

Date of Submission: Mar, 10, 2024 Date of Acceptance:

I. INTRODUCTION

The prevalence of online communication and transactions in the quickly developing information technology age made phishing attacks a growing concern to both individuals and organizations. Cybercriminals employ phishing, a deceitful technique, to manipulate users into disclosing sensitive information. Phishing puts the confidentiality and integrity of personal and organizational data at serious risk. The rising sophistication of approaches utilized by malevolent actors means that traditional security solutions frequently fall short of protecting against these dynamic threats. Therefore, the necessity for sophisticated and adaptable systems that can quickly recognise and stop phishing attacks is imperative. This work presents a complete method of phishing detection that combines behavioral analysis, domain-specific heuristics, and state-of-the-art machine learning techniques. By combining these components, our suggested system seeks to improve the precision and effectiveness of spotting phishing attempts on different websites, social media platforms, and emails.

Our suggested solution is intended to improve users' internet security by identifying and swiftly informing them of potential malware and phishing threats. This all-inclusive method includes Greek alphabet analysis, database comparison, port forwarding detection using Python tools, grammar analysis using NLP libraries, and many more. This paper follows the format Abstract, Introduction, Literature review, Methodology, User interface, result and discussion, benefits, Conclusion, Future Scope, and References.

II. LITERATURE REVIEW

Three methods were used for categorization by the authors of the paper [8], Rishikesh Mahajan and Irfan Siddavatam: Decision Tree, Random Forest, and Support Vector Machine. Their sample included 19,653 phishing URLs and 17,058 benign URLs, each with 16 attributes, collected from the websites of Alexa and PhishTank, respectively. The dataset was divided into testing and training sets in the following ratios: 50:50, 70:30, and 90:10, respectively.

The false positive rate, false negative rate, and accuracy score were among the metrics used to assess performance. They achieved 97.14% accuracy using the Random Forest algorithm as it has the lowest false negative rate. The study found that increasing the quantity of data used for training improves accuracy. Jitendra Kumar et al.'s study from [9] trained a range of classifiers, including K Nearest Neighbour, Random Forest, Naive Bayes Classifier, Logistic Regression, and Decision Trees, using variables extracted from the lexical structure of the URL. The URL dataset was created with volatility, overfitting, biased training, and unbalanced data in mind.

The dataset had an equal number of tagged phishing and genuine URLs and was further split into testing and training groups using a 7:3 ratio. Even though the AUC values of all the classifiers were almost equal, the Naive Bayes Classifier turned out to be more suitable because it had the highest value. The most accuracy that Naive Bayes could attain was 98%, with precision=1, recall=0.95, and F1-score= 0.97. Mehmet Korkmaz et al. developed a machine learning based phishing detection system in [10] by utilising eight different algorithms on three different datasets.

The techniques used were XGBoost, Random Forest (RF), Naive Bayes (NB), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbour (KNN), and Artificial Neural Network (ANN). It was discovered that the models using LR, SVM, and NB had low accuracy rates.

The techniques used were XGBoost, Random Forest (RF), Naive Bayes (NB), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbour (KNN), and Artificial Neural Network (ANN). It was discovered that the models using LR, SVM, and NB had low accuracy rates.

The 32-feature Kaggle dataset underwent principal component analysis (PCA) and various feature selection procedures. The dataset benefits from feature selection since it eliminates unnecessary or redundant data. The proposed model used the REF, Relief-F, IG, and GR algorithms for feature selection before turning to PCA. Random Forest achieved an accuracy rate of 97%.

It handled the over-fitting issue better and was less erratic. Using the UCI dataset, Abdulhamit Subasi et al. launched an intelligent phishing detection system in [12]. To identify phishing websites, a variety of machine learning techniques were employed as classifiers, including Artificial Neural Networks (ANN), K-Nearest Neighbours (K-NN), Support Vector Machines (SVM), C4.5 Decision Trees, Random Forests (RF), and Rotation Forests (RoF).

III. METHODOLOGY

A website that enables visitors to thoroughly examine the legitimacy of the URL using multiple layers of analysis. Greek Alphabet Analysis, Database Comparison, Grammar Analysis, and Port Forwarding Detection are among the procedures.

1. ANALYSIS OF GRAMMAR

Grammar analysis looks for irregularities or inconsistencies in the linguistic structure of textual information. Grammar analysis detects dishonest language patterns suggestive of fraudulent conduct, which improves the accuracy of phishing attempt detection.

- ***Input:*** A designated text box is filled with the URL to be evaluated.
- ***Process****:* Using preset rules and patterns, the algorithm thoroughly analyses the URL's grammar. This investigation looks for any irregularities or deviations that might point to a phishing effort or malevolent intent.
- ***Output:*** A result indicating whether or not the URL adheres to predicted grammatical patterns is produced by the grammar analysis.

2. DATABASE COMPARISION

Database comparison is the technique of comparing features and patterns between suspicious data and pre-existing phishing databases to find fraudulent information more rapidly. As a result, phishing detection systems are more successful overall.

- ***Input*****:** The URL under scrutiny is compared against a comprehensive database of known legitimate URLs and a database of known phishing URLs.
- ***Process*****:** The system utilizes predefined databases to check if the entered URL matches any known legitimate or phishing URLs. This database comparison helps identify URLs with a history of being associated with phishing attacks or legitimate entities.
- ***Output*****:** The result provides information on whether the URL matches any entries in the legitimate or phishing databases.

3. PORT FORWORDING DETECTION

Monitoring network traffic for instances of port forwarding is known as port forwarding detection, and it is an essential component of phishing detection since it can reveal any attempts to reroute users to malicious servers, strengthening defenses against phishing assaults.

- ***Input*****:** A phishing assault exploitable technique is detected by analyzing the URL.
- ***Process*****:** The system looks for port forwarding, which is the

process of diverting network traffic from one port to another. This method can be used by attackers to hide dangerous URLs.

***-Output*:** The result alerts the user to a possible security risk by indicating whether port forwarding is found in examined URL.

4. GREEK ALPHABET ANALYSIS

Greek alphabet analysis in phishing detection entails searching for Greek characters in the text since hackers commonly use them to circumvent standard filters. The ability of the system to identify and neutralize phishing threats is enhanced by locating these anomalies.

- ***Input*:** An analysis of the URL is performed with an emphasis on how the Greek alphabet is utilised in the URL structure.
- ***Procedure*:** The system looks for Greek letters in the URL because hackers could use them to make visually appealing URLs that trick visitors.
- ***Output*:** The result indicates if the URL contains any Greek alphabet characters, indicating a possible attempt to build a bogus URL.

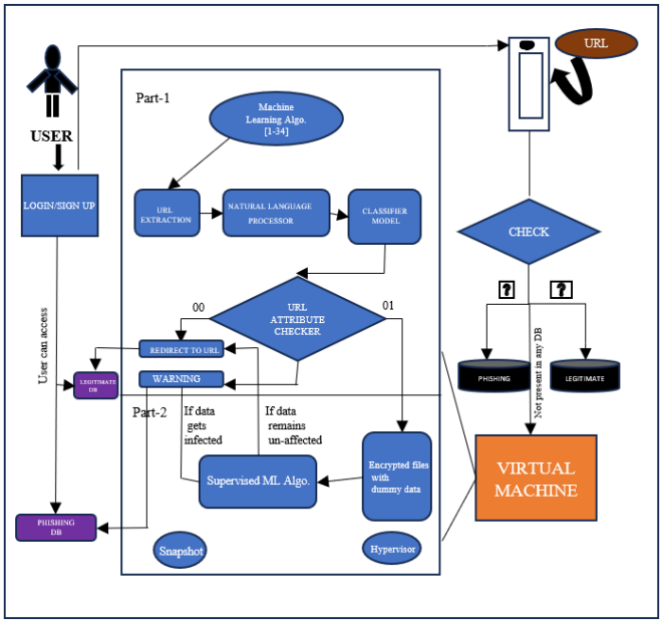

Fig.1- Architecture of Phishing Detection

5. USING IP ADDRESS

Verifying if the URL contains an IP address. Phishing websites frequently use direct IP addresses to conceal their genuine identities.

6. LONG URL

Recognizing unusually lengthy URLs, which are frequently used in phishing efforts to trick people.

7. SHORT URL

Examining abbreviated URLs, which are frequently used to hide their destination and can direct users to phishing websites.

8. SYMBOL@

Looking for the "@" symbol, as this could signal an attempt to trick consumers by imitating official email correspondence.

9. REDIRECTING//

Recognizing the usage of excessive redirects. It is a tactic that phishing websites frequently utilize to trick visitors.

10. PREFIX/ SUFFIX

Look closely at URL prefixes and suffixes, as phishers may use these to craft phony site addresses.

11. SUBDOMAINS

Counting the subdomains is important since an unusually high number could indicate a phishing effort.

12. HTTPS

Check if HTTPS is being used, as many phishing sites do not use secure connections, although genuine websites frequently do.

13. DOMAIN REGISTRATION LENGTH

Examining how long a domain has been registered, since shorter periods may be a sign of a fraudulent website.

14. FAVICON

Verifying the existence and authenticity of the favicon connected to the website, since phishers can trick users with phony favicons.

15. NON- STANDARD PORTS

Identifying non-standard ports used in URLs is important since it might be a sign of phishing activity. This is known as a Non-Standard Port.

16. HTTPS DOMAIN URL

Verifying that, while using HTTPS, the domain and the URL are consistent with each another, as differences could indicate phishing efforts.

17. REQUESTED URL

Examining the requested URL's validity because phishing websites may tamper with it to deceive users.

18. ANCHOR URL

Analyzing the anchor links' destination because they may lead to fraudulent websites.

19. LINKS IN SCRIPT TAGS

Finding links contained within script tags. It is a tactic that attackers frequently use to conceal harmful URLs.

20. SERVER FORM HANDLER

Examine the form handler on the server for any anomalies in form processing or other possible indications of phishing.

21. INFORMATION OF EMAIL

Recognizing shady email addresses which are connected to the website, since phishing websites could utilize phony contact details.

22. ABNORMAL URL

Recognizing URLs that don't follow standard patterns and indicating possible phishing attempts.

23. WEBSITE FORWARDING

Recognizing whether a website is sending users to another, since this may be a phishing technique.

24. STATUS BAR CUST

Examining changes made to the status bar that could be used to trick users about where a link leads.

25. DISABLED RIGHT CLICK

Tracking efforts to turn off the ability to right-click on web pages. It is a common tactic to keep users from using browser functions that would reveal phishing attempts.

26. USING POPUP WINDOW

Analyzing the use of popup windows, which phishing websites may employ to divert attention or deceive visitors.

27. IFRAME REDIRECTION

Determining whether iframe redirection—a method of sending viewers to phishing pages—is present.

28. AGE OF DOMAIN

Examining the domain's age is important because younger domains might be more connected to phishing.

29. DNS RECORDING

Verifying any irregularities in DNS recording, since phishers can alter DNS records to trick users.

30. WEBSITE TRAFFIC ANALYSIS

The traffic features of phishing websites frequently diverge from those of trustworthy websites.

31. PAGERANK

Determining a website's page rank; phishing sites may have artificially inflated or decreased ratings.

32. GOOGLE INDEX

Verifying whether Google has indexed the website since reputable websites have a higher chance of being listed.

33. LINKS POINTING TO PAGE

Analyzing how many external links lead to a particular website to determine its legitimacy or potential for phishing.

34. STATS REPORT

Analyzing the website's statistical reports and searching for any anomalies that might point to phishing activities.

35. CLASS

The class label that, after the previously specified components are looked at, indicates whether or not the situation is considered to be a phishing attempt.

IV. USER INTERFACE

User Interaction: The user hits the submit button to obtain the outcome after the analytic procedures.
- Output: The user sees a full result that summarizes the information gleaned from every level of investigation. This contains details on the URL's grammatical accuracy, database presence, port forwarding detection, and Greek character

usage.

To increase the identification of phishing attempts, this methodology offers a multidimensional approach to URL legitimacy assessment by integrating linguistic analysis, historical database comparison, and unique threat indicators.

V.RESULT & DISCUSSIONS

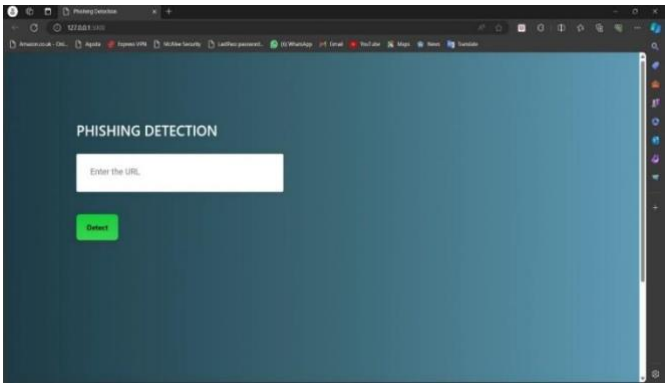

Fig.2 – User Interface

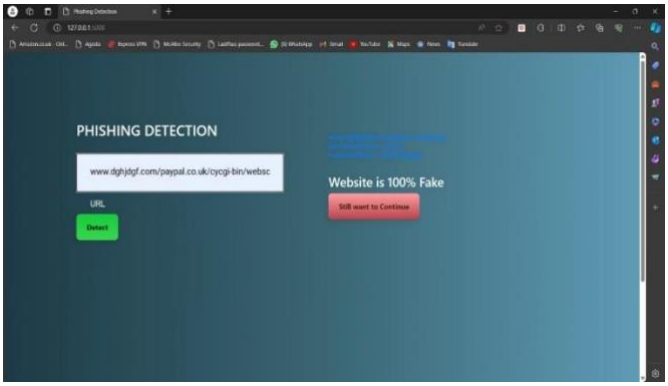

Fig.3- Detection of Phishing URL

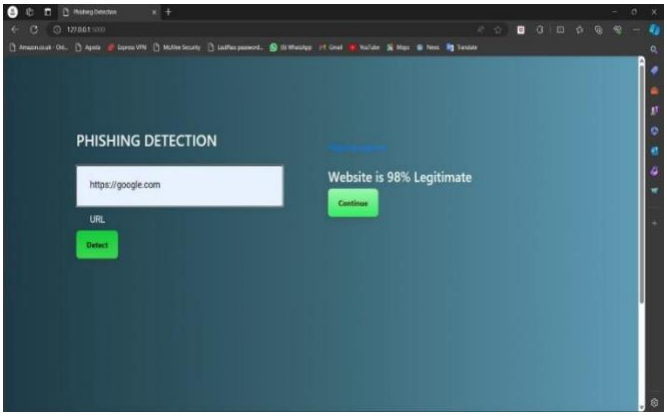

Fig.4- Detection of malicious URL

	ML Model	Accuracy	f1_score	Recall	Precision
0	Gradient Boosting Classifier	0.974	0.977	0.994	0.986
1	CatBoost Classifier	0.972	0.975	0.994	0.989
2	XGBoost Classifier	0.969	0.973	0.993	0.984
3	Multi-layer Perceptron	0.969	0.973	0.995	0.981
4	Random Forest	0.967	0.971	0.993	0.990
5	Support Vector Machine	0.964	0.968	0.980	0.965
6	Decision Tree	0.960	0.964	0.991	0.993
7	K-Nearest Neighbors	0.956	0.961	0.991	0.989
8	Logistic Regression	0.934	0.941	0.943	0.927
9	Naive Bayes Classifier	0.605	0.454	0.292	0.997

Fig.5-Accuracy of various models used for URL detection

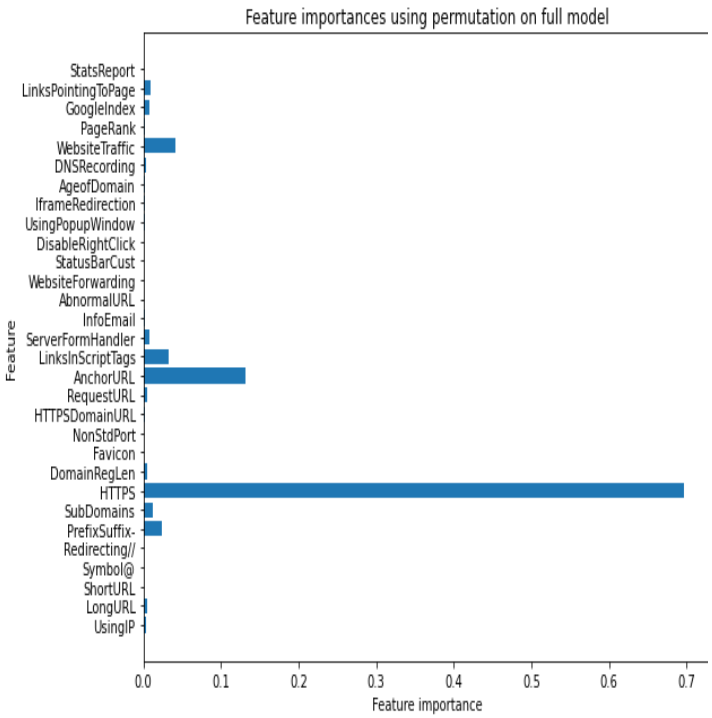

Fig.6- Feature importance for Phishing URL

VI. BENEFITS

These are the benefits for using a multi-layered approach for Phishing detection are

1) MORE ACCURATE DETECTION

***Smart Technology*:** We employ state-of-the-art technologies like machine learning to improve accuracy. This facilitates the system's gradual adaptability as it picks up new tricks for phishing.

Real-time Monitoring: The system keeps an eye on user activities in real-time, quickly spotting potential phishing threats and providing timely alerts.

2) DETECTING UNKNOWN ATTACKS

***Adaptive Systems*:** The system is designed to detect new and unknown phishing attacks by looking for unusual patterns in user behavior.

***Smart Guesswork*:** Heuristic methods help the system make educated guesses regarding potential threats, even if they're brand new or haven't been seen before.

3) ALERTING USERS

***Clear Warnings*:** When a threat is detected, the system sends clear and user-friendly warnings. These alerts help users understand the potential danger.
***User Preferences*:** Users can personalize their alert settings, ensuring that warnings are delivered in a way that suits their preferences.

4) MULTI-LAYERED APPROACH

***Many Checks*:** The system doesn't rely on a single method. It combines different ways of looking for phishing signs, including checking content, URLs, headers, and sender details.
Double-Check Strategy: By using both known patterns and analyzing behavior, the system becomes more reliable. This multi-layered approach minimizes mistakes and makes the system robust against different phishing tactics.

VII. CONCLUSION

In conclusion, our project represents a significant leap forward in online security by introducing a multi-layered approach to counter phishing and malware threats. By incorporating advanced techniques to detect semantic anomalies and staying vigilant against evolving cybercriminal strategies, our system ensures a robust defense mechanism. The prompt alerting of users adds an extra layer of protection, empowering individuals to make well briefed decisions in real time. With the overarching goal of substantially reducing success rates for both phishing and malware attacks, our contribution aims to create a safer digital environment. This innovative approach not only addresses current threats but also positions our system to adapt and evolve alongside the dynamic landscape of cyber threats, making a valuable contribution to the ongoing efforts to enhance online security.

VIII. FUTURE SCOPE

Phishing attacks continue to raise serious challenges to cybersecurity, necessitating continuous improvements in detection technologies and the incorporation of new functionality. In this work, a future scope for improving the security of Phishing Detection systems, with an emphasis on the development and implementation of unique features such as an Extension for E-mail and Message Applications, User Education initiatives, and an Authentication Page. These enhancements aim to strengthen existing detection techniques

while also addressing new attack pathways and user vulnerabilities. Our suggested paradigm stresses the convergence of technology breakthroughs and user-centric techniques to strengthen businesses' resilience to phishing threats.

IX. REFERENCES

[1] Anti-phishing Working Group (APWG) Phishing Activity Trends Report, 4th quarter 2020, https://docs.apwg.org/reports/apwg trends report q4 2020.pdf
[2] FBI Internet Crime Report 2020,

https://www.ic3.gov/Media/PDF/AnnualReport/2020 IC3Report.pdf

[3] Verizon 2020 Data Breach Investigation Report,

https://enterprise.verizon.com/resources/reports/2020-databreachinvestigations-report.pdf

[4] World Health Organization, Communicating for Health, Cyber Security, https://www.who.int/about/communications/cyber-security

[5] N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection-based associative classification data mining," Expert Systems with Applications, vol. 41, no. 13, pp. 5948–5959, 2014.

[6] K. L. Chiew, E. H. Chang, W. K. Tiong et al., "Utilization of website logos for phishing detection," Computers & Security, vol. 54, pp. 16–26, 2015.

[7] "Identifying fraudulent websites through fuzzy logic," K. M. Kumar and K. Alekhya, International Journal of Advanced Research in Computer Engineering Technology (IJARCET), 2016.

[8] Rishikesh Mahajan and Irfan Siddavatam, "Phishing website detection using machine learning algorithms," International Journal of Computer Applications (0975-8887), vol. 181, no. 23, 2018.

[9] Jitendra Kumar, A. Santhanavijayan, B. Janet, Balaji Rajendran, and Bindhumadhava BS, "Phishing website classification and detection using machine learning," International Conference on Computer Communication and Informatics (ICCCI), 2020

[10] Mehmet Korkmaz, Ozgur Koray Sahingoz, and Banu Diri, "Detection of phishing websites by using machine learning-based URL analysis," 11th International Conference on Computing, Communication, and Networking Technologies (ICCCNT), 2020

[11] Mohammad Nazmul Alam, Dhiman Sarma et al., "Phishing attacks detection using a machine learning approach," 3rd International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020

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

[13] Structure of a URL : image,

https://towardsdatascience.com/phishingdomain-detection-with-ml5be9c99293e5

[14] Rami M. Mohammad, Fadi Thabtah, and Lee McCluskey, "Phishing Website Features," University of Huddersfield, 2021.

[15] Arathi Krishna V*, Anusree A, Blessy Jose, Karthika Anilkumar and Ojus Thomas Lee, "Phishing Detection using Machine Learning based URL Analysis: A Survey", International Journal of Engineering Research & Technology (IJERT), 2021.