Proceedings of the 5th International Conference on Pervasive Computing and Social Networking (ICPCSN-2025)
DVD Part Number: CFP25EZ5-DVD; ISBN: 979-8-3315-3518-6

Securing the web using URL based analysis and Machine Learning Algorithms

Gattu Manik Sai
Artificial Intelligence and Data science
Vardhaman College of Engineering
Hyderabad, India
maniksai.gattu007@gmail.com

Chekka Mukesh Babu
Department of Artificial Intelligence and Data science Engineering
Vardhaman College of Engineering
Hyderabad, India
mukeshchekka@gmail.com

Y. Chandra Vamsi Karthik
Artificial Intelligence and Data science
Vardhaman College of Engineering
Hyderabad, India
chandravamsikarthik@gmail.com

Rajalingam S
Electrical and Electronics Engineering, Saveetha Engineering College, India
rajalingams@saveetha.ac.in

Hariharan Shanmugasundaram
Artificial Intelligence and Data Science
Vardhaman College of Engineering
Hyderabad, India
mailtos.hariharan@gmail.com

Karuppiah Natarajan
Electrical and Electonics Engineering, Vardhaman College of Engineering, Hyderabad,India
natarajankaruppiah@gmail.com

***Abstract*—As we have witnessed, the usage of the internet and smart phones has increased in recent years. This has caused a lot of fraud, scams, and the loss of personal information. Phishing attacks are the easiest way to get confidential information from users. This type of action is performed through phishing websites and various other platforms, including social media, e-commerce sites, etc. There are some websites that steal personal information, banking details, and important data like user IDs and passwords. All these are growing problems in the cyberworld that can affect the integrity of data, and they have to be resolved to achieve fairness in data. Phishing detection is a challenging problem, but many solutions were proposed, like rule-based analysis, list-based analysis, anomoly-based analysis, etc. So we have come up with a machine-learning-based URL analysis technique. This involves nine different algorithms and a dataset. The major algorithms will be gradient-boosting classifiers, multi-layer perceptrons, etc. We have implemented an interface that gives the output of the phishing probability of a given URL. The approach will depict that the proposed models have outstanding detection performance, accurate results, and a better success rate[1].**

***Index Terms*—Cybersecurity, Machine learning, Phishing, URL analysis, Detection, Feature Extraction**

I. Introduction

In the modern world, we deal with the internet and various digital platforms to get our work done. People use the internet and digital platforms in various fields, like education, banking, communication, tourism, shopping, etc. Today, businesses of all scales are using the internet and digital platforms for marketing, improving sales, storing customer details, etc. As the usage of the internet and mobile phones has increased, the availability of the internet everywhere at any time has also increased simultaneously. Access to the internet has enabled attackers to exploit the digital platforms from any location, which leads to the loss of sensitive information. This action will have a significant impact on the information security of organizations. To overcome cyberattacks, many safety measures have emerged and developed. As technology advances, organizations are implementing new methods and safety measures to protect users from such cyberattacks. Many experts have deployed advanced technologies to eradicate cyberattacks [2].

Cybercriminals, hackers, pirates, or attackers are the people who carry out cyberattacks. These unauthorized individuals aim to steal sensitive information from users in various ways. Since the attacks started in 1988, they have continuously been carried out on the internet up to the present day. These attacks mainly target multiple areas like fraud, scams, forgery, illegal information scraping, and more. These cybercrimes are increasing gradually, day by day. Cybercriminals are also involved in money laundering, providing fake credentials for criminals to attack legal websites, and illicit activities. Today, attackers carry out bank scams from anywhere by exploiting UPI and other online banking services. With the advancement of technology, attackers carry out insurance, finance, and investment banking scams. In addition to these, attackers have carried out numerous scams across various sectors. As more sectors join the online service for development, attackers are taking advantage of it to exploit the information[3].

Agreeing to the Phishing Action Report by the AntiPhishing Working Group (APWG), the full number of phishing websites watched within the to begin with quarter of 2022 surpassed one million. This surge has been especially apparent since the beginning of the COVID-19 widespread. These impacts are too apparent within the one of a kind brands being focused on by phishing campaigns, whose number has expanded altogether since the third quarter of 2020. Monetary administrations, which incorporate banks, are particularly prone to phishing. Within the to begin with quarter of 2024, this equipment seller segment was the foremost habitually victimized by phishing, with 23.6% of all assaults. Web and SaaS suppliers have moreover been focused on by a huge division of the assaults [5].

In figure.1, the bar graph indicates that phishing attacks are continuously rising from 2021 to 2024. The average weekly count of phishing attacks has increased by more than 90% since 2021. Today, this has become a serious issue for data privacy and protection. Despite the introduction of new techniques, attackers continue to find ways to steal your information through phishing attacks. So we can observe that the average attacks decreased from quarter 2 (Q2) to quarter 4 (Q4) in 2023. Once again, attacks increased by 20% in the first quarter (Q1) of 2024[6].

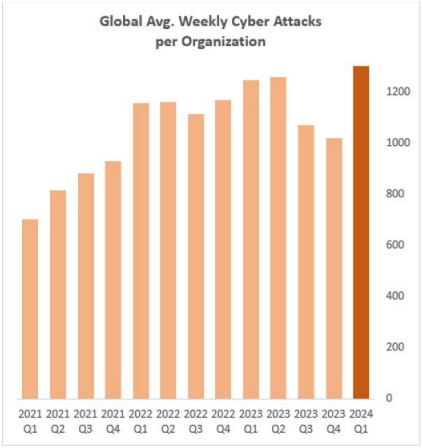

Fig. 1. Average weekly phishing attacks per organization from 2021-2024.

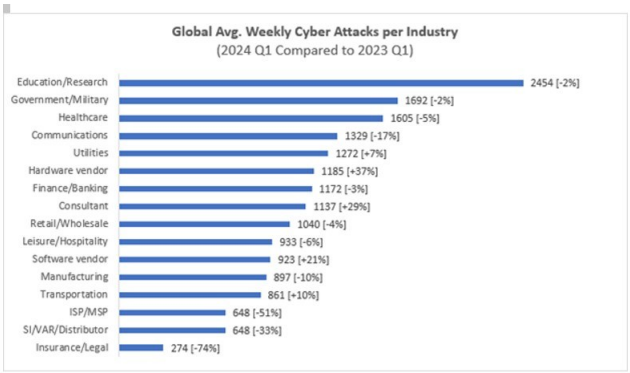

Fig. 2. Weekly cyber attacks per industry.

In figure.2, the bar graph shows a continuous rise in phishing attacks across all the sectors from 2023 to 2024, with some sectors showing an increase of more than 20%. This trend is likely to increase more in the upcoming days, so we have to find a solution to overcome the number of phishing attacks and increase security for the user.

This rise in phishing attacks indicates that attackers and hackers most commonly rely on them to steal information. They find it effortless to design and implement a duplicate copy of the website, as shown in Figure 3. The attackers create phishing websites, including a phishing kit. Subsequently, user clicks on the URL with the provided email. They think that the email is safe and secure. Once they click on this URL, they steal credentials and information from the user. Then they use this information and credentials to earn money from legitimate sources. The fake website looks similar to the legitimate site, but with minor differences that no one notices. In some situations, they create traffic on our network while accessing a legitimate site and start stealing information. They use other methods to extract information and steal money from the user. Some methods will include giving payment urls to click and asking users to share personal and bank details for verification[8].

In this way, phishing attacks take place. The attacker uses the stolen sensitive information and credentials to earn money. Previously, they used random emails to attack. Now, they prefer links that appear similar to reliable organization links. The user has to be careful while clicking on unusual links. In some situations, the attacker manipulates the user to click on the phishing links, which leads to data loss. The users may not always identify the attack by themselves. To overcome these attacks, the security experts implemented many detection techniques, such as Ensemble Learning, Deep Learning, Artificial Neural Networks, and Machine Learning. Previously, rule-based systems, list-based systems and similarity-based systems were also deployed. At present, machine learning models are in huge demand due to their ability to train, test, and have better accuracy. Although neural networks can also be implemented, they are more complex than machine learning models[9].

Therefore, this paper aims to develop a phishing detection system using Machine Learning Algorithms and feature extraction to analyze URLs. It also aims to implement a GUI interface to check whether the URL is phishing and help in checking the probability of an attack. The goal is to not rely on other websites or check lists to evaluate the URLs. The paper contains multiple sections: in the next section, the related works are included. In the third section, the details of the proposed work are outlined. In the fourth and fifth sections, results and discussions are shared, and conclusions are drawn, respectively[10].

II. LITERATURE SURVEY

A comprehensive literature review on the detection of phishing websites using machine learning reveals a dynamic landscape shaped by the evolving tactics of cyber criminals. Phishing attacks have steadily grown in sophistication, necessitating innovative approaches for their identification and mitigation. The combination of machine learning techniques with this domain has created significant attention. Researchers have explored various algorithms and models, including decision trees, support vector machines, neural networks, and ensemble methods, aiming to enhance the efficiency of phishing detection systems. These models leverage features derived from diverse sources,such as URLs, webpage content, SSL certificates, and user behavior data, to distinguish between legitimate and malicious websites.[3] Evaluation metrics like accuracy, precision, recall, and ROC curves have

become the yardsticks for assessing the effectiveness of these models. However, the literature also underscores the persistent challenges of dealing with imbalanced datasets, concept drift, and the relentless cat-and-mouse game between attackers and defenders. Recent advances in the field have witnessed the application of deep learning techniques, natural language processing, and the fusion of threat intelligence feeds, enabling more robust real-time detection capabilities. Case studies demonstrate the tangible benefits of these systems in real-world scenarios[11]. While significant progress has been made, the literature review reveals that there are still open challenges awaiting resolution. Future research directions include the development of more adaptive and explainable AI models, improved feature engineering methodologies, and strategies to counter emerging threats. In conclusion,the synthesis of existing knowledge in this literature review not only sheds light on the current state of machine learning-based phishing detection and it also underscores the importance of continuous innovation in this critical cyber security domain[12].

III. RELATED WORK

The purpose of this work was to analyze the webpage's URL in order to create a phishing detection system. A URL is a complicated string that contains syntactic and semantic expressions for an online resource. Upon closer inspection, Figure 4 displays the URL's structure. In the literature review, it was noted that useful attributes gleaned from the URL improve classification accuracy. Furthermore, characteristics like CSS, content, meta data, site structure, and the use of thirdparty services can all increase accuracy. The new websites that need to be classed will take longer to classify as a result of these features. It is anticipated that the suggested model, which was trained just using the URL's attributes, will categorize data faster than competing models. With this knowledge in mind, the study simply plans to analyze URLs. As a result, the classification outcomes of the acquired characteristics using various machine learning methods are contrasted. Additionally, the current results are contrasted with those of another study that used the same dataset[15]. Different URL addresses canbe created using fields like domain, subdomain, Top Level Domain (TLD), protocol, directory, file name, path, and query. In general, these associated fields in phishing URLs differ from those in authentic websites. As a result, URLs are crucial for identifying phishing attempts, particularly when it comes to rapidly categorizing web pages. Software Requirement Analysis for the Proposed System: Detection of Phishing Websites Using Machine Learning. An overview of our project's software requirements is provided below:

A. *Introduction:*

The proposed system aims to enhance the detection of phishing websites using machine learning techniques The system will encompass web data collection, feature extraction, model training, real-time detection, and reporting.

B. *Block Diagram*

The proposed system for detecting phishing websites using machine learning can be represented by a block diagram that consists of several interconnected components. At its core is the Machine Learning Model, which encompasses various machine learning models. These algorithms are responsible for analyzing features extracted from input data, such as URLs and webpage content. The Feature Extraction module pre-processes and transforms the input data into a format suitable for model analysis. A crucial element is the Threat Intelligence Integration, which provides real-time updates and data on emerging phishing threats. The Web-based Application serves as the user interface, facilitating user interactions with the system. User input, primarily URLs for analysis, is processed through the Client-Server Architecture, with the server hosting the Machine Learning Model and a Database for data storage. The system emphasizes security through Data Encryption, User Authentication, and User Privacy measures. It also ensures scalability via Load Balancing and caching mechanisms. Continuous Improvement is enabled through a Feedback Loop for user input and Model Retraining to adapt to evolving threats. Logging and Monitoring functionalities track system performance, and Reporting and Alerting mechanisms keep users informed of detection results. This block diagram illustrates the interplay of components in the proposed system, aiming to provide robust, real-time phishing detection with user-friendly access and a focus on security and adaptability.

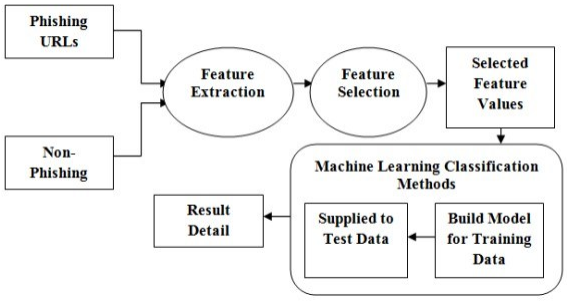

Fig. 3. Block Diagram.

C. *Activity Diagram*

An activity diagram is the way to clarify the stream of the method with the step by step when arrange of the method things. On the off chance that we begin with the primary stage of client side activity at that point client perform the header check for mail or URL check for web page he/she needs to visit some time recently giving any delicate data. The condition phishing check for the conceivable ways; on the off chance that YES implies not secure at that point in case you need more information at that point visit the wiki page for subtle elements or else exit. In case the condition NO suggests the secure surfing on the web. Enter detail, individual information and check sends are the activities can be taken by client and final organize gives the secure operation yield.

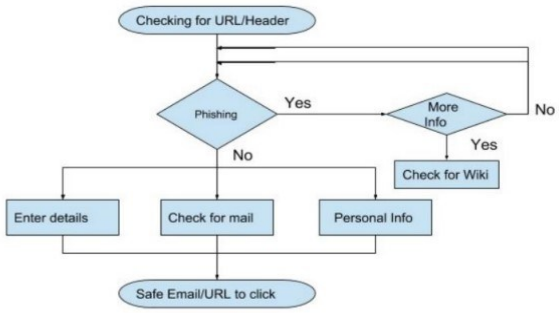

Fig. 4. Activity Diagram

D. *Use Case Diagrams*

The Use Case Diagram for the proposed system, which aims to detect phishing websites using machine learning, outlines the various interactions and functionalities involving different actors. The primary actors in this system include ”User,” ”Machine Learning Model,” and ”Threat Intelligence Service.” The ”User” actor interacts with the system through multiple use cases. Firstly, they can input a URL for analysis, triggering the ”Analyze URL” use case. This involves the system processing the URL through the Machine Learning Model for phishing detection. Users can also provide feedback on the detection results through the ”Provide Feedback” use case, contributing to system learning and improvement. The ”Machine Learning Model” actor encapsulates the core functionality of the system. It performs the analysis of input URLs, categorizing them as either phishing or legitimate. The model also supports periodic retraining using feedback data to enhance its accuracy. The ”Threat Intelligence Service” actor plays a vital role in keeping the system updated with real-time phishing threat information. It continuously feeds threat data into the system, ensuring that the Machine Learning Model is aware of the latest phishing tactics and trends. The Use Case Diagram illustrates how these actors interact to deliver a comprehensive phishing detection solution. Users can submit URLs for analysis, provide feedback, and benefit from an adaptive model, while the Threat Intelligence Service ensures that the system remains current and effective in countering evolving phishing threats.

E. *Testing*

- Data Splitting: The first step is to split the dataset into training, validation, and testing subsets. Typically, a significant portion of the data is used for training the machine learning models, while the validation set helps fine-tune hyperparameters and prevent overfitting. The testing set is kept separate and is used to evaluate the model’s performance objectively.
- Model Evaluation Metrics: To assess the model’s performance, various evaluation metrics are used, including True Positive Rate (TPR), False Positive Rate (FPR), accuracy, precision, recall, F1-score, and the Receiver Operating Characteristic (ROC) curve. These metrics provide insights into the model’s ability to correctly classify phishing websites while minimizing false positives.
- Cross-Validation: Cross-validation techniques such as kfold cross-validation can be employed to ensure that the model’s performance is consistent across different subsets of the data. This helps validate the model’s generalization capabilities.
- Hyperparameter Tuning: Fine-tuning of hyperparameters is performed to optimize the model’s performance further. Techniques like grid search or random search can be employed to find the best combination of hyperparameters.
- Real-time Testing: For practical deployment, the model should be tested in a real-time or near-real-time environment. It should continuously monitor incoming web requests or emails, classify them as either legitimate or phishing, and log the results for analysis.
- Security Testing: The system should be subjected to security testing to identify vulnerabilities and ensure that it cannot be exploited by attackers. This includes penetration testing and vulnerability assessments.
- Performance Testing: Performance testing ensures that the system can handle a high volume of web requests or emails without experiencing delays or failures. It assesses the system’s scalability and response time under various load conditions.
- Robustness Testing: Phishing attacks can take various forms, so the system should be tested against different types of phishing techniques, including spear-phishing and social engineering attacks.
- User Experience Testing: If the system interacts with users, usability and user experience testing should be conducted to ensure that it is user-friendly and intuitive. • Feedback Loop: Continuous monitoring and feedback from users and security experts are essential for refining the model and keeping it upto-date with emerging phishing threats. This feedback loop helps in adapting the model to evolving attack techniques.
- Deployment and Monitoring: Once the model passes all testing phases, it can be deployed in a production environment. Continuous monitoring and periodic retesting are essential to ensure that the model remains effective over time.

IV. RESULTS AND DISCUSSIONS

Table 1 presents the unsorted metrics and ML algorithms. Table 2 shows the results of various classifiers used for model training and various components of each classifier, which include accuracy, precision, and f-1 score. From the above table, we can find that the gradient boosting classifier is the best among all the 9 classifiers based on all the parameters. In the table 9 all the values are sorted based on accuracy where in table 1 all the values are unsorted.

Table1. Unsorted metrics of ML algorithms.

ML Model	Accuracy	f1 Score	Recall	Precision
Gradient Boosting Classifier	0.974	0.977	0.994	0.986
Random Forest	0.965	0.969	0.993	0.990
Support Vector Machine	0.964	0.968	0.980	0.965
Decision Tree	0.959	0.963	0.991	0.993
K-Nearest Neighbours	0.956	0.961	0.991	0.989
Logistic Regression	0.934	0.941	0.943	0.927
Naive Bayes Classifier	0.605	0.454	0.292	0.997
XGBoost Classifier	0.549	0.544	0.987	0.987
Multi-layer Perceptron	0.549	0.544	0.987	0.987

Table2. Sorted metrics of ML Algorithms.

ML Model	Accuracy	f1 score	Recall	Precision
Logistic Regression	0.934	0.941	0.943	0.927
K-Nearest Neighbours	0.956	0.961	0.991	0.989
Support Vector Machine	0.964	0.968	0.980	0.965
Naive Bayes Classifier	0.605	0.454	0.292	0.997
Decision Tree	0.959	0.963	0.991	0.993
Random Forest	0.965	0.969	0.993	0.990
Gradient Boosting Classifier	0.974	0.977	0.994	0.986
XGBoost Classifier	0.549	0.544	0.987	0.987
Multi-layer Perceptron	0.549	0.544	0.987	0.987

Fig. 5. The interface to enter the URL for Detection

Whenever the URL is given in the input block in figure 10, the input block takes an URL and gives the probability in the output block. In figure 11, the output screen shows the probability of a URL, whether it is legitimate or phishing, based on the analysis.

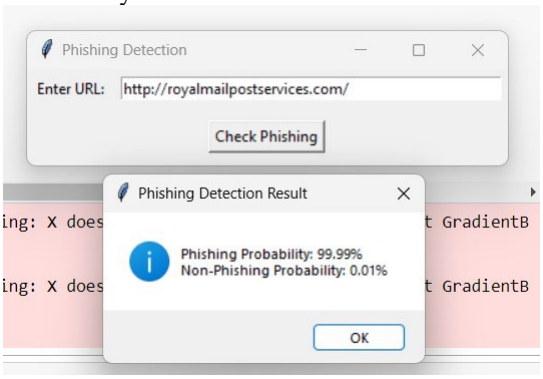

Fig. 6. The result of phishing probability after analysis of the URL

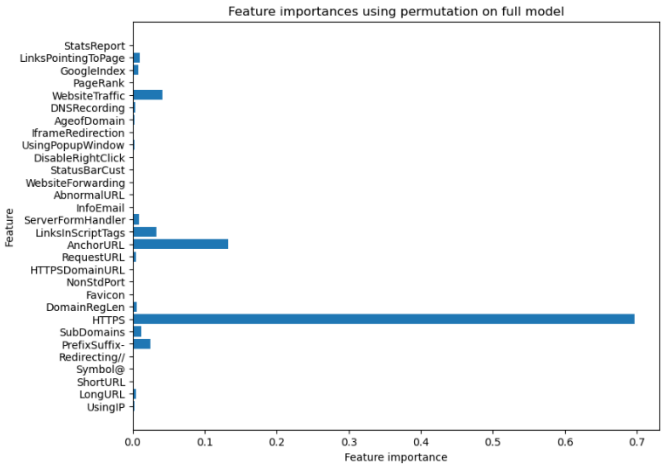

Fig. 7. Feature Importance using Permutation on full model

V. CONCLUSION

The conclusion from this extend is to investigate different machine learning models, perform Exploratory Information Examination on phishing dataset and understanding their highlights. Making this scratch pad made a difference me to memorize a part approximately the highlights influencing the models to distinguish whether URL is secure or not, moreover I came to know how to tuned demonstrate and how they influence the demonstrate execution. In conclusion, this project has successfully addressed the critical issue of phishing URL detection. Through meticulous data preprocessing, the development of a robust machine learning model, and rigorous evaluation, we have achieved commendable results in identifying malicious URLs. Despite encountering challenges in data collection and model optimization, the project's outcomes demonstrate its potential for enhancing online security. The lessons learned in the process, coupled with the extensive scope for future improvements outlined in the 'Future Scope' section, emphasize the significance of this project. By contributing to the fight against phishing attacks, this project holds promise in safeguarding users and organizations from cyber threats, underlining its relevance and impact in today's digital landscape. The ultimate conclusion on the Phishing dataset is that a few highlight like "HTTPS", "AnchorURL", "WebsiteTraffic" have more significance to classify URL is phishing URL or not. Slope Boosting Classifier accurately classify URL up to 97.4 particular classes and subsequently decreases the chance of noxious connections.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] G. Karatas, O. Demir and O. K. Sahingoz, "Deep Learning in Intrusion Detection Systems," 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), ANKARA, Turkey, 2018, pp. 113-116, doi: 10.1109/IBIGDELFT.2018.8625278.

[5] G. Karatas and O. K. Sahingoz, "Neural network based intrusion detection systems with different training functions," 2018 6th International Symposium on Digital Forensic and Security (ISDFS), Antalya, 2018, pp. 1-6, doi: 10.1109/ISDFS.2018.8355327.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] M. Babagoli, M. P. Aghababa, and V. Solouk, "Heuristic nonlinear regression strategy for detecting phishing websites," Soft Computing, vol. 23, no. 12, pp. 4315–4327, 2018.

[9] E. Buber, B. Diri, and O. K. Sahingoz, "Detecting phishing attacks from URL by using NLP techniques," 2017 International Conference on Computer Science and Engineering (UBMK), pp. 337-342, 2017.

[10] E. Buber, B. Diri, and O. K. Sahingoz, "NLP Based Phishing Attack Detection from URLs," Advances in Intelligent Systems and Computing Intelligent Systems Design and Applications, pp. 608–618, 2018.

[11] R. M. Mohammad, F. Thabtah, and L. Mccluskey, "Predicting phishing websites based on self-structuring neural network," Neural Computing and Applications, vol. 25, no. 2, pp. 443–458, 2013.

[12] A. K. Jain and B. B. Gupta, "Towards detection of phishing websites on client-side using machine learning based approach," Telecommunication Systems, vol. 68, no. 4, pp. 687–700, 2017.

[13] F. Feng, Q. Zhou, Z. Shen, X. Yang, L. Han & J. Wang, "The application of a novel neural network in the detection of phishing websites," Journal of Ambient Intelligence and Humanized Computing, pp 1-15, 2018.

[14] S. Smadi, N. Aslam, and L. Zhang, "Detection of online phishing email using dynamic evolving neural network based on reinforcement learning," Decision Support Systems, vol. 107, pp. 88–102, 2018.

[15] R. S. Rao and A. R. Pais, "Detection of phishing websites using an efficient feature-based machine learning framework," Neural Computing and Applications, vol. 31, no. 8, pp. 3851–3873, Jun. 2018.

[16] T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," 2018 IEEE 12th International Conference on Semantic Computing (ICSC), pp. 300-301, 2018.

[17] R. S. Rao, T. Vaishnavi, and A. R. Pais, "CatchPhish: detection of phishing websites by inspecting URLs," Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 2, pp. 813–825, Oct. 2019.

[18] PhishTank-Friends of PhishTank," PhishTank [Online].Available:https://www.phishtank.com/friends.php. [Accessed: 09-Mar-2020].