ISSN: 2395-7852

International Journal of Advanced Research in Arts, Science, Engineering & Management

Volume 10, Issue 3, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 6.551

+91 9940572462 +91 9940572462 ijarasem@gmail.com www.ijarasem.com

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 6.551 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 10, Issue 3, May 2023 |

Detection of Phishing Website Using Machine Learning

Shivani [1], Shraddha D S[2], Shraddha S Shetty[3], Shravya S[4] ,Deepak M Rao[5]

Student, Department of Computer Science and Engineering, Shri Madhwa Vadiraja Institute of Technology and Management, Bantakal, Karnataka, India[1,2,3,4]

Asst. Professor, Department of Computer Science and Engineering, Shri Madhwa Vadiraja Institute of Technology and Management, Bantakal, Karnataka, India[5]

ABSTRACT: Phishing attacks, a prevalent threat to online security, employ fake websites to steal sensitive information from naive users. These phishing URLs might be challenging to spot since attackers commonly use evasion techniques like URL obfuscation and redirection. In this study, we propose a machine learning-based technique for detecting phishing URLs. We look at a group of URLs and pull out a number of features, including features based on domains, lexicons, and content. In order to determine whether a URL is phishing or not, we train decision trees, random forests, and neural networks, among other machine learning methods. Our experiments show that the proposed technique successfully recognises phishing URLs with high rates of accuracy, with the neural network model achieving a 98.5% accuracy. The findings highlight the potential of this strategy as a practical tool for increasing internet security and demonstrate how machine learning algorithms are quite good at spotting phishing URLs. The suggested method offers a reliable way to identify phishing websites and might be added to current cyber security programmes to increase their use.

KEYWORDS - Phishing, Classification model, Cyber Crime, Machine Learning Algorithm**.**

I.INTRODUCTION

Security researchers are currently particularly concerned about the issue of phishing. This is due to how easy it is to make a fake website that mimics the actual one. Few people possess the skills necessary to recognise bogus websites, leaving them open to phishing assaults. Experts can recognise fake websites. Phishing is a type of internet fraud that involves claiming to be a reliable website in order to obtain consumers' sensitive or private information. Phishing attacks employ a number of strategies, including website spoofing, social engineering, manipulating links, and covert redirects. The attackers' main goal is to gain access to bank accounts. The cost to US businesses of phishing victims being their clients is $2 billion a year. The third "Microsoft Computing Safer Index Report," which was published in February 2014, estimates that lost productivity due to phishing may cost $5 billion annually worldwide. Attacks involving phishing are more successful because consumers aren't aware of them. Phishing assaults are extremely difficult to fight since they prey on user vulnerabilities, so it is imperative to advance phishing detection techniques. Updating Internet Protocol (IP) URLs that have been blacklisted in antivirus databases is a common step in the "blacklist" technique, which is used to identify phishing websites. Attackers employ obfuscation and other straightforward techniques to get around blacklists by altering URLs to look legitimate to people who can be duped. There are numerous methods to eliminate phishing websites, Tools that can be used with each of these methods include network security, authentications, client-side tools, user education, server-side filters, and classifiers. at various points in the attack flow. a heuristic-based detection with a very high false positive rate that takes into consideration characteristics that are occasionally but not always present in real phishing attempts. The drawbacks of the blacklist and heuristic-based approaches are currently being addressed by a large number of security researchers employing machine learning techniques. In order to forecast future behaviour or events, many algorithms used in machine learning technology rely on historical data. The programme examines a number of valid URLs that have been blocked using this method.

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 6.551 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 10, Issue 3, May 2023 |

II. LITERATURE REVIEW

Altyeb Taha, proposed a Intelligent Ensemble Learning approach For Phishing Website Detection based on Weighted Soft Voting [1].

To outperform a single classifier, ensemble learning combines the predictions of multiple different classifiers. In order to improve the identification of phishing websites, this research suggests an intelligent ensemble learning strategy based on weighted soft voting. The websites were first classified as phishing or authentic websites using a base classifier made up of four heterogeneous machine-learning algorithms. A novel weighted soft voting technique based on Kappa statistics is the second. According to the experimental findings, the suggested intelligent strategy for phishing website identification surpassed basic classifiers and soft voting methods and achieved the greatest accuracy of 95% and an Area Under the Curve (AUC) of 98.8%. Since there is no one-size-fits-all strategy for phishing elimination due to the constantly evolving nature of these attacks, more efficient and improved ways for identifying them are needed. The following are the study's main contributions and the significance of each for enhancing phishing website detection: In order to boost the detection accuracy of phishing websites, a novel weighted soft voting method based on the k-statistic is developed. This method evaluates the contributions of each classifier and assigns greater influence weights to stronger classifiers and lower impact weights to weaker classifiers.

S.T.Deepa, Dr.K.K. Thanammal, Dr.S.S. Sujatha, Phishing Website Detection Using Novel Features And Machine Learning Approach [2].

Phishing is a type of digital attack that negatively impacts people by tricking the victim into divulging their private and sensitive information, such as passwords for accounts, bank information, ATM pin-card information, and so forth. Consequently, protecting sensitive data from malware or web phishing is difficult. The goal of this project is to identify phishing websites utilizing cutting-edge characteristics and machine learning techniques. Utilizing a convolutional auto encoder, the input URL websites are first feature extracted. These features are then passed to a deep neural network classifier for improved classification of real and phishing URLs. The two categories of phishing sites have real repercussions for Internet users and organizations, such damaging brand value and raising customer agitation rates. Models to distinguish phishing attempts based on ordering outdated pages can be created using ML techniques, and the programme can then use these models. The basic methodology evaluated different aspects of a URL, the next methodology dissected the authority of a site and determined if the site was presented and it also looked at who was in charge of it, and the third methodology evaluating the validity of the site looked into AI strategies and evaluated their performances when ready to be applied to datasets containing features that can distinguish between a Phishing Website and a secure one. Each phishing site was associated with an original, one-of-a-kind finger impression made from the combination of suggested highlights. Then, based on Convolutional auto encoder, the feature collector extracts features. The features that were extracted, including those based on the address bar, domains, HTML, and JavaScript.

Shwetha, Prof. Kavitha, Detection of Phishing Website using Machine Learning [3].

In order to compare and forecast more accurately, this research provides a framework for phishing identification that uses different machine learning algorithms, such as logistic regression, random forest, and support vector machines. It also covers data visualization, data analysis, and phishing website detection. Researchers have used another method in addition to monitoring website traffic on Alexa to identify fraudulent websites. At that time, the coordinating computation determines the URL score. This site is flagged as a phishing site if this score is higher than a specified edge esteem. Each URL is shown as a binary feature vector. Pre-processing, data interpretation, data visualization, and determining if a URL leads to a legitimate website or a phishing website are all included in the work-recommended framework. used three machine learning techniques to distinguish between real and phishing websites: LR, RF, and SVM. Getting the dataset ready. Based on the data set of the model being trained, the data set is provided to the machine learning model. Following study, a model prediction based on an inference drawn from training data sets is made. A phishing URL can be recognized by comparing its design to a legitimate URL.

No.	Paper Title	Author Name	Key Points	Remark
1	Intelligent Ensemble Learning approach For Phishing Website Detection based on Weighted Soft Voting	Altyeb Taha, 2021	Proposes an intelligent ensemble learning approach for phishing website detection based on weighted soft voting to enhance the detection of phishing websites [1].	The experimental results showed that the suggested intelligent approach for phishing website detection outperformed the base classifiers and soft voting method and achieved the highest accuracy of 95% and an Area Under the Curve (AUC) of 98.8%.
2	Phishing Website Detection Using Novel Features And Machine Learning Approach	S.T.Deepa,Dr.K.K. Thanammal, Dr.S.S.Sujatha, 2021	1) Aims at detecting phishing website using novel features and machine learning algorithm. 2)Used three machine learning algorithms (LR), Random Forest (RF), vector support (SVM) to identify websites as legitimate and phishing [2]	Shows that the implemented system is best in detecting phishing websites with 89% accuracy.
3	Detection of Phishing Website using Machine Learning	Shwetha, Prof. Kavitha S.N, 2020.	Used list based method, later used heuristic approach. This uses database of signature for any known attacks that matches the signature of the heuristic template to check the legitimate or phishing website[3]	Prevent the user from malware attack and have recommended to use a layer of security.

In summary, the work presented in this paper is built on previous research to explore how security of data stored on cloud relates to people's trust. While earlier work focused on data storage impacts people, we focus on its impact on the world wide acceptance of cloud.

III. METHODOLOGY OF PROPOSED SURVEY

The process for identifying phishing URLs entails a number of procedures, including::

- User sees a website: During this stage, the user accesses what they think is a phishing URL on a website.
- User features are then collected from the webpage by conducting an analysis on it, such as the user's IP address, browser type, and operating system.
- Data processing: The retrieved characteristics that have been preprocessed are then ready for analysis. It may be necessary to perform cleaning, normalization, and feature engineering.
- Dataset and Feature Extraction: The machine learning model is trained using a set of well-known authentic and phishing URLs. The length of the URL, the age of the domain, the existence of particular keywords, and other indicators of phishing activity may all be variables in the model.
- Machine Learning Classification: A machine learning model is trained on the extracted attributes to discover trends in the data and classify URLs as real or phishing.
- Flask API: Once the model is complete, it can be incorporated into a Flask API to respond to user requests and deliver classification outcomes in real time.
- Prediction: After obtaining the preprocessed information, a machine learning classification model leverages patterns found from a labelled dataset to assess whether a URL is real or phishing.
- Phishing or Legitimate: The result of the prediction is then conveyed to the user using the Flask API, indicating whether the URL is a phishing site or not.

This technology combines site scraping, data processing, machine learning, and API development to provide an automated system for detecting phishing URLs. By looking at a range of user and URL-related data, the system can accurately identify potential threats and provide users with the information they need to stay secure online. Figure 1

shows the flow diagram for detecting phishing websites using machine learning algorithms.

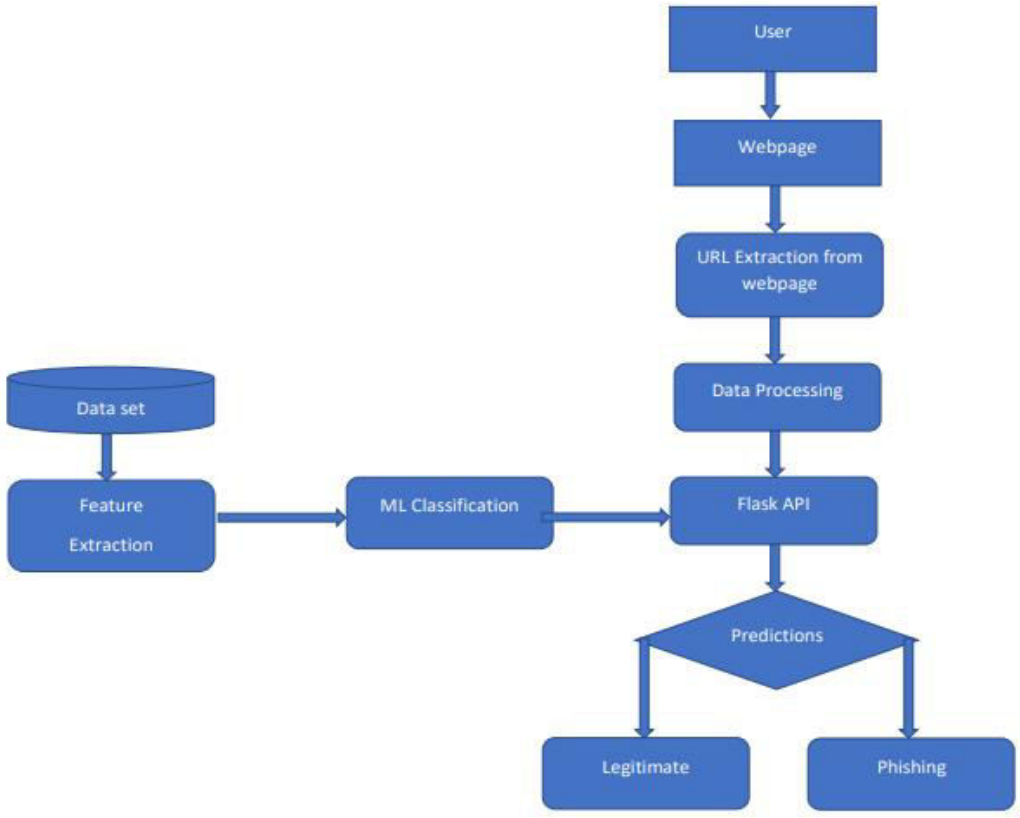


Figure 1: Flow Diagram

The algorithms used to detect phishing websites include Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Naïve Bayes Classifier, Decision Tree, Random Forest, Gradient Boosting Classifier, CatBoost Classifier, and Multi-layer Perceptron. Figures 2, 3, 4, and 5 show the results of testing the accuracy of these algorithms..

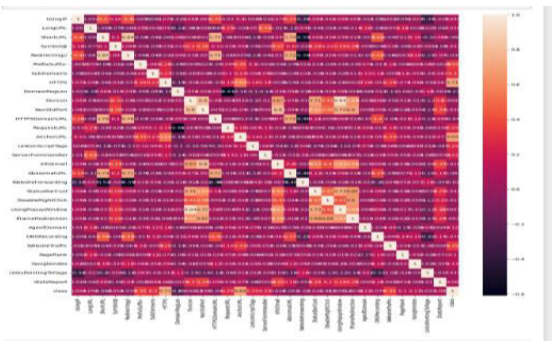


Figure 2

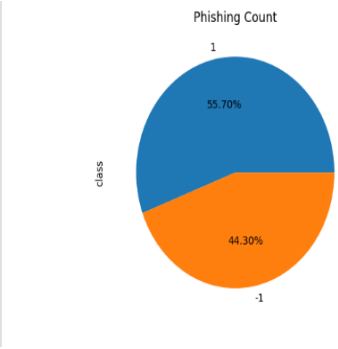


Figure 3

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 6.551 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 10, Issue 3, May 2023 |

	precision	recall	f1-score	support
-1	0.99	0.96	0.97	976
1	0.97	0.99	0.98	1235
accuracy			0.97	2211
macro avg	0.98	0.97	0.97	2211
weighted avg	0.97	0.97	0.97	2211

Figure 4

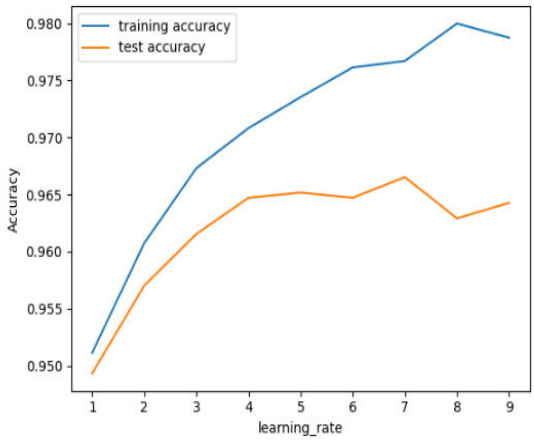


Figure 5

IV. CONCLUSION AND FUTURE WORK

Machine learning techniques may make it possible to identify phishing websites. Numerous studies have shown that these algorithms can detect fake websites with high accuracy rates by looking at factors including website content, domain names, and user behaviour.

The fact that these algorithms might not be perfect and produce false positives or false negatives must always be kept in mind. It is suggested that these algorithms be used in conjunction with other anti-phishing strategies, such as user education and awareness initiatives, in order to create a comprehensive strategy for defeating phishing efforts.

The only line of defence against phishing attacks should not therefore rely solely on machine learning algorithms, even though they may be a useful tool for spotting phishing websites. Using a multi-layered strategy that combines technological solutions with user awareness and education is the best way to prevent phishing attacks and mitigate their effects.
In the current digital environment, phishing websites are becoming more and more common since they may be used to steal sensitive data, such login passwords or financial information. Future improvements, like the following, may help to better identify these websites: Information exchange and cooperation Collaboration between businesses and individuals can enhance the ability to detect new phishing assaults.

In conclusion, leveraging cutting-edge technology and interdisciplinary cooperation will be necessary for future improvements in phishing website detection in order to increase the precision of this detection and shield people from phishing scams.

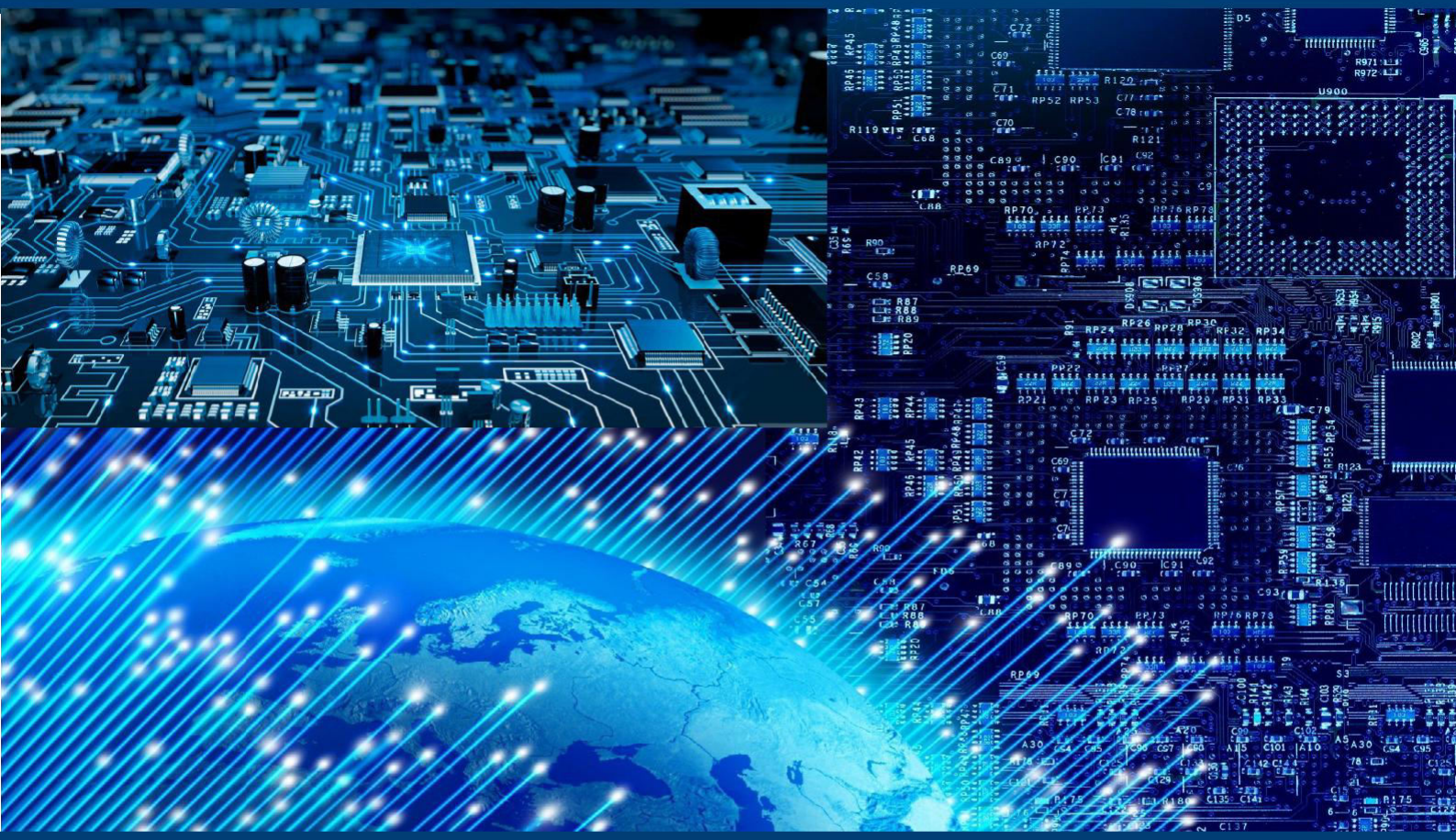
REFERENCES

[1] Altyeb Taha, “Intelligent Ensemble Learning approach For Phishing Website Detection based on Weighted Soft Voting” , 2021.
[2] S.T. Deepa, K.K. Thanammal, S.S. Sujatha, “Phishing Website Detection using Novel Features And Machine Learning Approach” in Turkish Journal of Computer and Mathematics Education 2021..
[3] Shwetha, Prof. Kavitha S.N, “Detection of Phishing Website using Machine Learning”, 2020.
[4] Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh and Dr. Aram Alsedrani, “Detecting Phishing Websites using Machine Learning” in IEEE 2019.
[5Ammar Odeh, Ismail Keshta and Eman Abdelfattah, “Machine Learning Techniques for Detection of Website Phishing: A Review for Promises and Challenges” in IEEE 2021.

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 6.551 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 10, Issue 3, May 2023 |

[6] Moitrayee Chatterjee and Akbar Siami Namin, "Detecting Phishing Websites through Deep Reinforcement Learning" in IEEE 43rd Annual Computer Software and Applications Conference 2019.
[7] Ammar Odeh, Ismail Keshta and Eman Abdelfattah, "Machine Learning Techniques for Detection of Website Phishing: A Review for Promises and Challenges" in IEEE 2021.
[8] Vayansky, Ike, and Sathish Kumar. "Phishing-challenges and solutions." Computer Fraud & Security" 2018.1 (2018):15-20.
[9] Vaibhav Patil, Pritesh Thakkar, Chirag Shah, Tushar Bhat and S.P. Godse, "Detection and Prevention of Phishing Websites using Machine Learning Approach" in IEEE 2018.
[10] Malaika Rastogi, Anmol Chhetri, Divyanshu Kumar Singh, Gokul Rajan V, "Survey on Detection and Prevention of Phishing Websites using Machine Learning" in International Conference on Advance Computing and Innovative Technologies in Engineering 2021.



International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| Mobile No: +91-9940572462 | Whatsapp: +91-9940572462 | ijarasem@gmail.com |