ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 04, April 2025

Phishing Website Detection Using Deep Learning

[1] Pradeep Nayak, [2] Fathima Thahiba, [3] Shramik S Shetty, [4] Akash K Acharya, [5] Vandan M Shetty

[1] [2] [3] [4] [5] Dept of Information Science and Engineering, Alva's Institute of Engineering and Technology, Mangalore, India
Emails ID: [1] pradeep@aiet.org.in, [2] fathimathahiba@gmail.com

***Abstract*— Phishing attacks continue to be a concern, in the realm of security as they target sensitive information by using deceptive websites. The usual methods of detecting phishing websites heavily rely on heuristics and manual inspection. These approaches have limitations when it comes to adapting to phishing tactics. In our study we put forward an approach for identifying phishing websites using learning techniques. We utilize a neural network architecture that analyzes both the content and structure of websites. This architecture allows us to automatically detect phishing websites by learning patterns that indicate behavior. Additionally we explore the application of networks to capture time based changes in website content. This further enhances our models ability to identify phishing attacks that evolve over time. To assess the effectiveness of our approach we conduct experiments on datasets and compare the performance of our deep learning model with traditional machine learning methods. Our results clearly demonstrate that our proposed deep learning approach surpasses existing techniques in terms of accuracy and reliability when it comes to detecting phishing websites across scenarios. The findings from this study carry implications for enhancing security as they offer a more efficient and scalable solution, for detecting phishing websites. By utilizing deep learning methods we can improve the effectiveness of security systems, in detecting and minimizing the threats presented by phishing attacks. This in turn ensures the protection of users sensitive data, within our growing environment.**

Index Terms*— *Phishing, Cyber security, Cyber-attack detection, Deep Learning.

I. INTRODUCTION

Internet use has become an essential part of our daily activities as a result of rapidly growing technology. Due to this rapid growth of technology and intensive use of digital systems, data security of these systems has gained great importance. The primary objective of maintaining security in information technologies is to ensure that necessary precautions are taken against threats and dangers likely to be faced by users during the use of these technologies. The word 'Phishing' initially emerged in 1990s. The early hackers often use 'ph' to replace 'f' to produce new words in the hacker's community, since they usually hack by phones.

Phishing is a new word produced from 'fishing', it refers to the act that the attacker allure users to visit a faked Website by sending them faked e-mails (or instant messages), and stealthily get victim's personal information such as user name, password, and national security ID, etc. These information then can be used for future target advertisements or even identity theft attacks (e.g., transfer money from victims' bank account). The frequently used attack method is to send e-mails to potential victims, which seemed to be sent by banks, online organizations, or ISPs. In these e-mails, they will make up some causes, e.g. the password of your credit card had been misentered for many times, or they are providing upgrading services, to allure you visit their Web site to conform or modify your account number and password through the hyperlink provided in the e-mail. You will then be linked to a counterfeited Web site after clicking those links. The style, the functions performed, sometimes even the URL of these faked Web sites are similar to the real Web site. It's very difficult for you to know that you are actually visiting a malicious site. If you input the account number and password, the attackers then successfully collect the information at the server side, and is able to perform their next step actions with that information (e.g., withdraw money out from your account). Phishing itself is not a new concept, but it's increasingly used by phishers to steal user information and perform business crime in recent years. Within one to two years, the number of phishing attacks increased dramatically. According to Gartner Inc., for the 12 months ending April 2004, "there were1.8 million phishing attack victims, and the fraud incurred by phishing victims totaled $1.2 billion".

To address this challenge, we propose a novel approach that uses deep learning techniques for phishing website detection in order to overcome this difficulty. Our method focuses on using the Random Forest algorithm and Support Vector Classifier to extract features from URLs. These features are then fed into a CNLSTM architecture for classification. This enables us to identify intricate relationships and patterns within URLs that may point to the intention of phishing.

A. Proposed Scope and Contributions

This research paper's suggested scope is to examine and show how well deep learning methods combined with URL feature extraction can identify phishing websites. The Random Forest technique and Support Vector Classifier will be used to extract features from URLs, and then the CNLSTM architecture will be used for classification. By utilizing cutting-edge machine learning and deep learning techniques, the research seeks to improve the precision and resilience of phishing website identification while addressing the shortcomings of conventional phishing detection methods. The scope of the suggested technique encompasses

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 04, April 2025

an experimental evaluation that compares its performance with traditional methods, utilizing a dataset of legitimate and known phishing websites. Contributions:

Novel Approach: By combining deep learning methods with URL feature extraction, this study suggests a revolutionary method for detecting phishing websites. With the help of the CNLSTM architecture for classification and the Random Forest algorithm and Support Vector Classifier for feature extraction, it is possible to distinguish phishing websites from legitimate ones using only the features of their URLs.

Enhanced Robustness and Accuracy: The suggested strategy seeks to outperform more conventional approaches in terms of phishing website detection accuracy and robustness by utilizing deep learning techniques. The model can recognize intricate patterns and relationships inside URLs thanks to the integration of feature extraction algorithms with an advanced deep learning architecture, which improves the accuracy of phishing website detection.

Experimental Validation: Using a large dataset of well-known phishing and trustworthy websites, an experimental evaluation is part of the research. The outcomes of the experiment illustrate how well the suggested strategy works to identify phishing websites, indicating that it has the potential to surpass existing approaches and improve cybersecurity technologies.

Insights and Implications: This study's conclusions offer important new information about how to use deep learning methods to identify phishing websites. The factors that have been discovered as contributing to the classification can provide important insights into the properties of phishing URLs. These insights can then be leveraged to improve cybersecurity protocols and create more potent defenses against phishing assaults.

Future Directions: The suggested study creates opportunities for additional machine learning and cybersecurity research. It lays the groundwork for future research into deep learning architectures, feature engineering methods, and the incorporation of new data sources to improve phishing detection systems' detection capabilities.

II. RELATED WORK

A. Conventional Methods:

Heuristic-Based Detection: To find possible phishing websites, many early phishing detection systems relied on heuristic criteria. Criteria like mismatched URLs, dubious domains, or recognized phishing indications were frequently added in these filters. Heuristic-based strategies are straightforward and simple to use, but their capacity to adjust to novel phishing tactics may be constrained.

Blacklist-Based Detection: These systems keep track of well-known phishing domains and URLs, which they then compare against incoming requests in a database. These systems can be vulnerable to false positives and have trouble with zero- day assaults, even while they work well against recognized threats.

B. Methodologies Based on Machine Learning:

Supervised Learning: The detection of phishing websites has been approached using a variety of supervised learning methods, including logistic regression, decision trees, and support vector machines. These algorithms can identify intricate patterns in URL characteristics, although they do require labeled training data.

Unsupervised Learning: Anomalies in URL structures that might point to phishing activity have been found using unsupervised learning techniques like clustering algorithms. These methods may need more powerful processing power, but they can be helpful in spotting phishing trends that haven't been seen before.

C. Deep Learning-Oriented Strategies:

Neural Network designs: Phishing website detection has been accomplished through the use of deep learning designs such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and their variations. In comparison to conventional machine learning methods, these models have the ability to automatically learn hierarchical representations of URL features, which may allow them to capture more subtle patterns. *Transfer Learning:* Transfer learning methodshave demonstrated potential in enhancing performance, particularly in situations when labeled training data is scarce. They include fine- tuning a pre-trained deep learning model on a phishing detection task.

D. Extracting Features from URLs:

Lexical analysis: In this process, characteristics of URLs are extracted, such as the length of the domain, the existence of subdomains, and lexical attributes (such as the presence of hyphens or numbers). These characteristics can be utilized to create phishing detection classifiers that work well.

Behavioral Analysis: In addition to URL features, some methods concentrate on examining user behavior, such as mouse movements or click patterns. These techniques try to increase the detection accuracy of phishing attempts by merging user behavior with URL analysis.

III. ALGORITHM USED

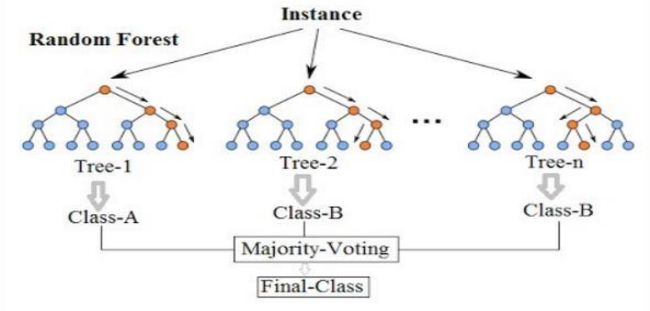
A. Random Forest Algorithm:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set.

ISSN (Online) 2394-2320

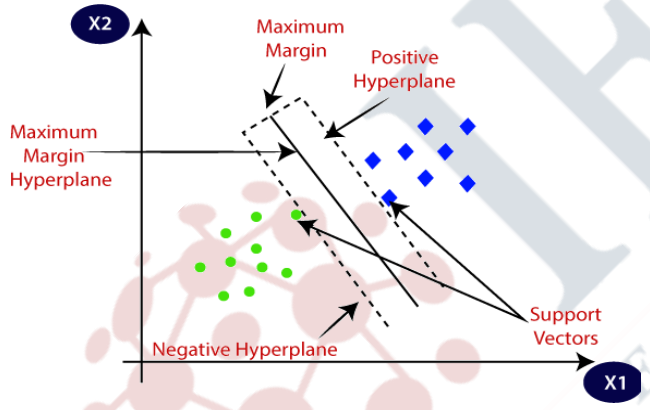
International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 04, April 2025



B. Support Vector Classifier:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n- dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.SVM chooses the extreme points/vectors that help in creating the hyper plane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyper plane.



IV. EXPERIMENTAL RESULT AND ANALYSIS

We used TensorFlow and Keras, two well-known deep learning tools, along with Python to accomplish our suggested strategy. In order to extract pertinent characteristics from the URLs, the Random Forest algorithm and Support Vector Classifier were used in the URL feature extraction phase. The CNLSTM architecture was then trained using these features as an input for classification. We divided the dataset according to a conventional train-test split, using 80% for training and 20% for testing. We used metrics like recall, accuracy, precision, and F1 score to assess our model's performance.

Analysis: Our proposed approach's effectiveness is validated by the high levels of accuracy, precision, recall, and F1 score that we achieved from our studies. Our model can detect phishing websites with strong accuracy because it combines the CNLSTM architecture with deep learning approaches for URL feature extraction to identify intricate patterns and relationships inside URLs. Important input for the CNLSTM architecture for classification is efficiently extracted from the URLs by the Random Forest algorithm and Support Vector Classifier.

Additionally, the feature importance analysis showed that a few URL features—like the age of the domain, its length, and the existence of subdomains—had a major impact on the classification results. This knowledge can help to better understand the traits of phishing URLs and improve the feature selection procedure in subsequent model iterations.

V. CONCLUSION AND FUTURE WORK

The proliferation of online transactions and purchases has been greatly aided by technological advancements, which simplify our daily lives. However, when sensitive information is exchanged online, it might result in unlawful access to the data of persons, businesses, or consumers. The most crucial component in defending consumers from information theft by phishers during online communication is security. One established method of obtaining user information is phishing, which uses a URL that appears exactly like the webpage in question. One important step in stopping hackers from accessing user data is identifying phishing attacks. Given that the number of victims is increasing due mostly to ineffective adoption of security technologies, users must be protected from cyberattacks with a clever strategy. Deep learning has demonstrated to be a beneficial advancement in comparison to conventional signature-based and classic machine learning-based solutions because of its high performance and end-to-end problem-solving capabilities, despite the rapid development of deep learning approaches.

The LSTM, CNN, and LSTM–CNN algorithms were presented in this work to identify and categorize website URLs as either authentic or phishing. The evaluation of the suggested approach showed that phishing website detection produced outstanding results. The performance of the suggested deep learning methods on the same dataset varied. In terms of accuracy, the CNN algorithm fared better than LSTM–CNN and LSTM, reaching 99.2%, compared to 97.6% and 96.8% for LSTM–CNN and LSTM, respectively. In order to validate the states of websites and increase the overall accuracy of training procedures, our future goals include shortening training times and optimizing feature engineering. Additionally, we plan to introduce a method that takes the webpage context and URL in order to detect phishing websites.

REFERENCES

[1] Ludl, C., McAllister, S., Kirda, E., & Kruegel, C. (2007, July). On the effectiveness of techniques to

detect phishing sites. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment* (pp. 20-39). Springer, Berlin, Heidelberg

[2] Zhang, Y., Hong, J., & Cranor, L. CANTINA: A Content- Based Approach to Detect Phishing Web Sites. In *Proceedings of the 16th International Conference on World Wide Web, Banff, AB, Canada* (pp. 639-64S).

[3] Pan, Y., & Ding, X. (2006, December). Anomaly based web phishing page detection. In *2006 22nd Annual Computer Security Applications Conference (ACSAC'06)* (pp. 381392). IEEE.

[4] Xiang, G., & Hong, J. I. (2009, April). A hybrid phish detection approach by identity discovery and keywords retrieval. In *Proceedings of the 18th international conference on World Wide Web* (pp. 571-580).

[5] Ebbu2017 Phishing Dataset. Accessed 24 July 2018. https://github.com/ebbubekirbbr/ pdd/ tree/master/ input.

[6] https://en.wikipedia.org/wiki/Precision_and_recall

[7] Buber, E., Diri, B., & Sahingoz, O. K. (2017b). NLP based phishing attack detection from URLs. In A. Abraham, P. K. Muhuri, A. K. Muda, & N. Gandhi (Eds.), *Intelligent systems design and Applications, springer international Publishing, cham (pp. 608–618).*

[8] Machine learning based phishing detection from URLs Ozgur Koray Sahingoz a, Ebubekir Buber b, Onder Demir b, B Diri - Expert Systems with Applications, 2019 – Elsevier

[9] Peng, T., Harris, I., & Sawa, Y. (2018). Detecting phishing attacks using natural language processing and machine learning. In IEEE 12th international conference on semantic computing (ICSC) (pp. 300–301)

[10] Adebowale, M.; Lwin, K.; Hossain, M. Intelligent phishing detection scheme using deep learning algorithms. *J. Enterp. Inf. Manag.* **2020**. [**Google Scholar**] [**CrossRef**]

[11] Zhang, L.; Zhang, P. PhishTrim: Fast and adaptive phishing detection based on deep representation learning. In Proceedings of the 2020 IEEE International Conference on Web Services (ICWS), Beijing, China, 19–23 October 2020; pp. 176–180. [**Google Scholar**] [**CrossRef**]

[12] Janet, B.; Reddy, S. Anti-phishing System using LSTM and CNN. In Proceedings of the 2020 IEEE International Conference for Innovation in Technology (INOCON), Bangaluru, India, 6–8 November 2020; pp. 1–5. [**Google Scholar**] [**CrossRef**]

[13] URL 2016|Datasets|Research|Canadian Institute for Cybersecurity|UNB. Unb.ca. 2022. Available online: **https://www.unb.ca/cic/datasets/url-2016.html** (accessed on 28 November 2020).

[14] Mahdavifar, S.; Ghorbani, A. Application of deep learning to cybersecurity: A survey. *Neurocomputing* **2019**, *347*, 149–176. [**Google Scholar**] [**CrossRef**]

[15] Chai, J.; Zeng, H.; Li, A.; Ngai, E.W.T. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Mach. Learn. Appl.* **2021**, *6*, 100134. [**Google Scholar**] [**CrossRef**]

[16] Adebowale, M.A.; Lwin, K.T.; Hossain, M.A. Deep Learning with Convolutional Neural Network and Long Short-Term Memory for Phishing Detection. In Proceedings of the 2019 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), Island of Ulkulhas, Maldives, 26– 28 August 2019; pp. 1–8. [**Google Scholar**] [**CrossRef**]

[17] Bahnsen, A.C.; Bohorquez, E.C.; Villegas, S.; Vargas, J.; González, F.A. Classifying phishing URLs using recurrent neural networks. In Proceedings of the 2017 APWG Symposium on Electronic Crime Research (eCrime), Phoenix, AZ, USA, 25–27 April 2017; pp. 1–8. [**Google Scholar**] [**CrossRef**]

[18] Chen, W.; Zhang, W.; Su, Y. Phishing detection research based on LSTM recurrent neural network. In *International Conference of Pioneering Computer Scientists, Engineers and Educators*; ICPCSEE 2018: Zhengzhou, China, 2018; pp. 638–645. [**Google Scholar**]

[19] Ariyadasa, S.; Fernando, S.; Fernando, S. Detecting phishing attacks using a combined model of LSTM and CNN. *Int. J. Adv. Appl. Sci.* **2020**, *7*, 56–67. [**Google Scholar**]

[20] Pham, T.; Hoang, V.; Ha, T. Exploring Efficiency of Character-level Convolution Neuron Network and Long Short Term Memory on Malicious URL Detection. In Proceedings of the 2018 VII International Conference on Network, Communication and Computing–ICNCC 2018, Taipei City, Taiwan, 14–16 December 2018. [**Google Scholar**]

[21] Lakshmi, V.; Vijaya, M. Efficient prediction of phishing websites using supervised learning algorithms. *Procedia Eng.* **2012**, *30*, 798–805. [**Google Scholar**] [**CrossRef**][**Green Version**]

[22] Malicious Url Recognition and Detection Using Attention- Based CNN-LSTM-KSII Transactions on Internet and Information Systems (TIIS)|Korea Science. Available online: **https://www.koreascience.or.kr/article/JAKO201905959996575.page** (accessed on 20 June 2022).

[23] Zhang, Q.; Bu, Y.; Chen, B.; Zhang, S.; Lu, X. Research on phishing webpage detection technology based on CNN-BiLSTM algorithm. *J. Phys. Conf. Ser.* **2021**, *1738*, 012131. [**Google Scholar**] [**CrossRef**]

[24] Jawade, J.V.; Ghosh, S.N. Phishing website detection using Fast.ai Library. In Proceedings of the 2021 International Conference on Communication information and Computing Technology (ICCICT), Mumbai, India, 25–27 June 2021. [**Google Scholar**] [**CrossRef**]