International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 12 August 2023 **Revised:** 10 October 2023 **Accepted:** 22 October 2023*

An Analysis of Malicious URL Detection Using Deep Learning

Maruti Patil[1], Dr.Sangram Patil[2]
[1]PhD Scholar, Computer Science & Engineering
D.Y. Patil Agriculture & Technical University,Talsande
Kolhapur,India
Patilmaruti16@gmail.com
[2]Associate Professor, Computer Science & Engineering
D.Y. Patil Agriculture & Technical University,Talsande
Kolhapur,India
sangrampatil@dyp-atu.org

Abstract— Considerable progress has been achieved in the digital domain, particularly in the online realm where a multitude of activities are being conducted. Cyberattacks, particularly malicious URLs, have emerged as a serious security risk, deceiving users into compromising their systems and resulting in annual losses of billions of dollars. Website security is essential. It is critical to quickly identify dangerous or bad URLs. Blacklists and shallow learning are two techniques that are being investigated in response to the threat posed by malicious URLs and phishing efforts. Historically, blacklists have been used to accomplish this. Techniques based on blacklists have limitations because they can't detect malicious URLs that have newly generated. In order to overcome these challenges, recent research has focused on applying machine learning and deep learning techniques. By automatically discovering complex patterns and representations from unstructured data, deep learning has become a potent tool for recognizing and reducing these risks. The goal of this paper is to present a thorough analysis and structural comprehension of Deep Learning based malware detection systems. The literature review that covers different facets of this subject, like feature representation and algorithm design, is found and examined. Moreover, a precise explanation of the role of deep learning in detecting dangerous URLs is provided.

Keywords- Malicious web sites, Cyberattacks, phishing,Deep Learning

I. INTRODUCTION

Millions of services are made available to people all over the world through the World Wide Web and the technologies that support it. We rely on websites for everything from basic information searches to paying for our purchases. As the Internet evolves and expands, more and more of our activities—including e-commerce, business, social networking, and banking—are now carried out online, increasing the risk of online crime. Consequently, protecting the internet is becoming more and more crucial. The importance of cyber security is rising in today's digital environment. The emergence of numerous types of major cyber-attacks has been linked to the expansion of Internet usage and the development of network technology. Every 39 seconds, there is an occurrence of a web attack somewhere in the world. Cybercriminals now use advanced methods to assault users due to the World Wide Web's growing prominence. These assaults include malware installation on user PCs, phishing and other forms of financial fraud, and shady websites advertising fake items. Designing reliable systems to identify cyber breaches is difficult due to the variety of attack tactics, including hacking attempts, drive-by downloads, SQL injections, and more. Due to the rapid changes in IT technology, the scarcity of security personnel, and the exponential emergence of new security risks, traditional security management platforms have limits. One of the most serious concerns is the production and dissemination of malicious Uniform Resource Locators (URLs).

There are two primary methods for identifying rogue URLs. The first method is a blacklist-based strategy, in which a database of known harmful URLs is updated regularly. This approach is used by well-known commercial systems like Google Safe Browsing, McAfee Site Advisor, and Web sense Threat Seeker. It is successful in recognizing known dangerous URLs but fails to pick up on fresh, previously undiscovered hazardous URLs that constitute a serious threat to consumers. As a result, these systems might not be entirely prepared to shield people from new online threats. The second method is heuristic-based. With better generalization capabilities based on traits or behaviours, it develops and retains signatures of recognized attacks in a blacklist. This approach has drawbacks since attackers can adapt their attacks to avoid detection. This method leaves systems open to previously unknown dangers and renders them incapable of properly identifying new hazardous URLs, jeopardizing user security and safety. The second method involves identifying hazardous URLs using

Artificial Intelligence (AI) methods, specifically Machine Learning (ML) classification models. Due to the fact that it addresses the shortcomings of the previous technique, this strategy has grown in favor over the past ten years. A training set of both harmful and benign URLs is necessary, and numerous attributes connected to the URLs are gathered in order to execute ML-based detection. Once the ML model has been trained on this dataset, it may use what it has learned to reliably classify new URLs and detect those that might be hazardous.

II. LITERATURE REVIEW

Machine learning is used in the Hung Le, Quang Pham, Doyen Sahoo, and Steven C.H. Hoi [1] approach to detect dangerous URLs. Before employing machine learning models like SVMs, the most popular and scalable techniques extract Bag-of-words-like features from the lexical properties of the URL string. There are more elements that experts created to enhance the model's prediction capabilities using Deep Learning. Using the proposed URLNet framework, an end-to-end deep learning system, a nonlinear URL embedding for harmful URL detection straight from the URL is constructed. A convolutional neural network is used to train URL embeddings in an optimal framework, specifically for the characters and words that make up the URL string. This method overcomes the limitations of the pre-existing models and enables the model to capture a variety of semantic information types. To address the issue of the excessive number of uncommon terms found in this work, advanced word-embeddings are also suggested. The research highlights a few shortcomings of current methods for detecting fraudulent URLs, but it makes no specific reference of any shortcomings of its suggested URLNet architecture. However, the following are some possible restrictions on the suggested framework: The framework may not perform well on URLs that are significantly different from those in the training data as it may not be able to capture all possible variations in URL structures and semantics. For the suggested methodology to effectively learn the URL embeddings, a substantial volume of training data is needed. Since the system relies on character and word embedding's that are trained on English text, it might not be able to handle URLs in languages other than English.

Tomas RASYMAS, Laurynas DOVYDAITIS [2] analysed on how to identify phishing URLs using machine learning approaches. Many features related to classification are covered in the review, such as word and character level embeddings, third-party features, lexical features, and third-party features. Additionally, the article looks at several neural network topologies—recurrent neural networks, convolutional neural networks, and a combination of the two—to help classify phishing URLs. In contrast to previous research, the paper compares its own method of merging various features and employing deep neural network architecture. The research does not investigate additional potential features that can enhance the model's accuracy; instead, it solely concentrates on three categories of features: lexical, character-level, and word

Bronjon Gogoi, Tasiruddin Ahmed, Arabinda Dutta[3] strategy that combines established blocklist techniques with cutting-edge machine learning (ML) strategies for identifying dangerous URLs. The hybrid strategy tries to improve URL detection by combining the best features of both approaches. The conventional method is a blocklist- or signature-based method that can identify existing harmful URLs. A massive dataset with 2.5 million benign and malicious URLs is used to train and assess the machine learning and deep learning-based method. Precision, recall, and the f1-score are all more than 0.97 in the system that combines deep learning, shallow learning, and traditional blocklist techniques. The suggested system employs CNN, LSTM, and CNN-LSTM as its deep learning models.

The ISCX-URL-2016 dataset, which consists of more than 110,000 URLs, was used by the Emine UCAR, Murat UCAR, Mürsel Ozan İNCETAS[4] to evaluate the effectiveness of their detection system. To differentiate between good and bad URLs, they used deep learning techniques like recurrent and convolutional neural networks. According to the experimental findings, the CNN model detects malicious URLs with a good accuracy rate.

Animesh Bhagwat, Kuldeep Lodhi, Shreyas Dalvi, Umesh Kulkarni [5] reviewed a summary of earlier research into using machine learning algorithms to detect potentially dangerous URLs. Among the significant research that the article cites are:

The study by Frank Vanhoenshoven, which used machine learning as the most effective method for tackling the binary classification problem of detecting dangerous URLs.

The study by Abu-Nimeh et al., which evaluated six different classifiers and determined that Random Forest had the lowest mistake rate.

The work by H.B. Kazemian and S. Ahmed, which compared three supervised ML models to two unsupervised ML models and discovered that supervised learning techniques were better appropriate in this situation.

The study by Amruta R. Nagaonkar and Umesh L. Kulkarni, which suggested five different techniques for identifying fraudulent URLs, including host-based and lexical aspects.

The study by Ahmed Abbasi and Hsinchun Chen, which evaluated the system design, accuracy, previous findings, and detection rates of various fraudulent website detection technologies.

The research suggests a supervised learning strategy for machine learning to identify dangerous URLs. The paper employs the following techniques:

A thorough analysis of methods for the discovery of dangerous websites is done.

A prediction model for identifying fake URLs is constructed by contrasting various supervised machine learning methods, such as Random Forest, K-Nearest Neighbor, Support Vector Machine, Decision Tree, and Artificial Neural Network.
The classifier model is trained using a dataset containing a large number of characteristics, and the most important and influential features are chosen to be part of the model.
An add-on for Google Chrome was developed to provide information on whether a website is harmful or not.

Arijit Das, Ankita Das, Anisha Datta, Shukrity Si and Subhas Barman[5] analysed that URLs can be successfully classified as dangerous or benign using deep learning models, more specifically the CNN-LSTM architecture. Deep learning methods solve the drawback of the hard-coded characteristics used in previous efforts by learning from patterns found in such URLs to extract features of their own. The CNN-LSTM design surpasses the simple RNN, simple LSTM, and simple LSTM architectures in a comparative study, with an accuracy of 93.59%. The researcher also examines the study's shortcomings and offers ideas for future research trajectories.

Farhan Douksieh Abdi and Lian Wenjuan [7]practise the simple method of Convolutional Neural Network (CNN) to ascertain whether a URL is safe or not. This algorithm's performance is compared in the paper to that of the Support Vector Machine (SVM) and the Logistic Regression (LR) techniques. The tests performed on 344821 benign URLs and 75643 malicious URLs reveal that the proposed algorithm detects malicious URLs with an accuracy rate of more than 96%. The research finds that the suggested approach can increase the generality of malicious URL detectors and is quick and accurate in identifying new malicious content.

A CyberLen deep learning-based approach is proposed by Yunji Liang, Qiushi Wang, Kang Xiong, Xiaolong Zheng, and Zhiwen Yu [8] to reliably and accurately identify harmful URLs. The system uses a factorization engine (FM) to explore latent interactions between lexical properties and a temporal convolutional network (TCN) to explore long-range correlations between URLs. To lessen the uncertainty of URL tokens, position embedding is used. To train a robust model, an effective wide- and deep-learning technique is suggested. The testing findings demonstrate the superior performance of the suggested approach for the reliable and effective detection of dangerous URLs. A limitation of the proposed system is that its performance has only been evaluated on one dataset; on other datasets, it might exhibit dissimilarities in performance.

Using machine learning and deep learning models, Clayton Johnson, Bishal Khadka, Ram B. Basnet, and Tenzin Doleck[9] look into the identification and categorization of risky URLs. In comparison to common machine learning techniques like Random Forest, CART, and kNN, the study assesses how well-known deep learning framework models like Fast.ai and Keras-TensorFlow perform on CPU, GPU, and TPU architectures. The researcher's conclusions indicate that firms intending to create URL filter applications or those looking to utilize machine learning to improve their present ones should adopt the Random Forest model. According to the study, some lexical characteristics found in URLs can be used to lower a deployed model's overhead expenses.

The Yuchen Liang, Xiaodan Yan[6] suggests a method for identifying malicious URLs, which have the potential to seriously compromise network security, based on the Deep Bidirectional LSTM model. For comparison, the research also discusses conventional machine learning techniques that categorize harmful websites using lexical features. The results demonstrate that the DBLSTM classifier performs far better than traditional machine learning methods, with a 98.6% accuracy rate. The paper finds that the suggested technique can be utilized to improve network security in the energy internet domain and is effective a[7]t identifying bad URLs.

Table 1: A comparison of various methods for identifying malicious URLs

Sr No	Paper Title	Technique used	Conclusion
1	URLNet: Learning a URL Representation with Deep Learning for Malicious URL Detection[1]	Convolutional Neural Networks (CNN) are used to train URL embedding in an optimized system and extract features similar to words in bags of words from URL strings.Utilizing cutting-edge word-embedding methods to address the issue of unusual words, which is typically seen in malicious URL detection tasks	The performance of the various URLNet variations, including character-level, word-level, and complete, is comparable, with URLNet(complete) routinely outperforming the others. While character-level URLNet is more efficient at higher FPRs, word-level URLNet performs better at low FPRs. Both types' advantages are combined in URLNet(Full).
2	Detection of Phishing URLs	It uses a branching neural network architecture	Give 94.4% accurate result.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 12 August 2023 **Revised:** 10 October 2023 **Accepted:** 22 October 2023*

	by Using Deep Learning Approach and Multiple Features Combinations [8]	with several hidden layers.	
3	A Hybrid approach combining blocklists, machine learning and deep learning for detection of malicious URLs[3]	a strategy that combines deep learning (DL), machine learning (ML), and traditional blocklist techniques. TensorFlow and Keras frameworks are used to build deep learning models, the sklearn framework is used to implement deep ML models.	The system is trained and tested 2.5 million malicious and benign URLs, and It accomplishes recall, precision, and a f1-score of greater than 0.97.
4	A DEEP LEARNING APPROACH FOR DETECTION OF MALICIOUS URLS[4]	Deep learning model CNN, RNN used to detect harmful URLs,The ISCX-URL-2016 dataset is used.	The CNN model gives 98% accuracy for binary classification and 95% for multi-class classification for the detecting dangerous URLs.
5	An Implemention of a Mechanism for Malicious URLs Detection[9]	This technique employs a cross-platform Google Chrome extension: JavaScript in the extension forwards user-entered URLs to Python for feature extraction. A serialized supervised learning model evaluates gathered	Random Forest is chosen as the classifier for the classification of harmful or benign URLs.
		features to indicate the site's maliciousness level.	
6	Deep Approaches on Malicious URL Classification [5]	For the purpose of classifying harmful URLs, the article employs three distinct architectures: simple RNN, simple LSTM, and CNN-LSTM.	With a 93.59% accuracy rate, the CNN-LSTM architecture trumps the other two. The Adam optimizer was used by the authors to train their model over 120 iterations at a learning rate of 0.0001.
7	Deep learning methods for malicious URL detection using embedding techniques as Logistic Regression with Lasso penalty and Random Forest[10]	Especially in the NLP method, the time frequency-inverse document frequency (TF-IDF) vector quantifier uses N-gram parameters for feature extraction, in this study it is proposed based on deep learning to find the path of a bad URL.	The DNN model, which includes logistic regression and L1 (Lasso) penalty as feature selection, gives the best results with 96.95% accuracy, 99% precision, 100% recall and 99% F1 score.
8	Robust Detection of Malicious URLs With Self-Paced Wide & Deep Learning[11]	The research suggests a CyberLen deep learning-based method for reliably and effectively identifying dangerous URLs. The system uses a temporal convolution network and a factorization machine (FM).	In comparison to all baselines, the FM-TCN-SPLD solution performs better with a 5% performance margin.

9	Towards Detecting and Classifying Malicious URLs Using Deep Learning[12]	Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) are two deep learning models employed in the study that produced very accurate results in the identification and classification of harmful URLs.	Based on restrictions related to time, performance, and complexity, Random Forest was deemed to be the best model.
10	Using Deep Learning to Detect Malicious URLs[6]	For the purpose of identifying malicious URLs, the paper's primary approach is a Deep Bidirectional LSTM (DBLSTM) model. There are four interacting neural network layers, including gates that add or remove information from the cell state, in each LSTM cell that makes up the DBLSTM layers.	The DBLSTM classifier outperforms traditional machine learning methods with an accuracy rate of 98.6%.

III. CONCLUSION.

We have investigated many methods for identifying malicious URLs. With the use of deep learning techniques, we examined the feasibility of accurately identifying dangerous URLs based on the results of this survey. There are several models and methodologies under investigation, and each has benefits and drawbacks. In general, deep learning models, especially those that mix CNN and LSTM architectures, have the potential to improve URL detection systems. However, the performance of these models can vary depending on the dataset and specific use case. Future research may focus on removing the restrictions and enhancing the accuracy and generalizability of these models.

REFERENCES

[1] H. Le, Q. Pham, D. Sahoo, and S. C. H. Hoi, “URLNet: Learning a URL Representation with Deep Learning for Malicious URL Detection,” Feb. 2018, [Online]. Available: http://arxiv.org/abs/1802.03162

[2] L. A. R and S. Thomas, “DETECTING MALICIOUS URLS USING MACHINE LEARNING TECHNIQUES: A COMPARATIVE LITERATURE REVIEW,” International Research Journal of Engineering and Technology, vol. 269, 2008, [Online]. Available: www.irjet.net

[3] B. Gogoi, T. Ahmed, and A. Dutta, “A Hybrid approach combining blocklists, machine learning and deep learning for detection of malicious URLs,” in Proceedings - 3rd IEEE India Council International Subsections Conference: Impactful Innovations for Benefits of Society and Industry, INDISCON 2022, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/INDISCON54605.2022.9862909.

[4] E. Uçar, “A DEEP LEARNING APPROACH FOR DETECTION OF MALICIOUS URLS.” [Online]. Available: https://www.researchgate.net/publication/338477987

[5] A. Bhagwat, S. Dalvi, K. Lodhi, and U. Kulkarni, An Implemention of a Mechanism for Malicious URLs Detection.

[6] A. Das, A. Das, A. Datta, S. Si, and S. Barman, “Deep Approaches on Malicious URL Classification.”

[7] F. Douksieh Abdi and L. Wenjuan, “Malicious URL Detection Using Convolutional Neural Network,” International Journal of Computer Science, Engineering and Information Technology, vol. 7, no. 6, pp. 01–08, Dec. 2017, doi: 10.5121/ijcseit.2017.7601.

[8] Y. Liang, Q. Wang, K. Xiong, X. Zheng, Z. Yu, and D. Zeng, “Robust Detection of Malicious URLs with Self-Paced Wide & Deep Learning,” IEEE Trans Dependable Secure Comput, vol. 19, no. 2, pp. 717–730, 2022, doi: 10.1109/TDSC.2021.3121388.

[9] C. Johnson, B. Khadka, R. B. Basnet, and T. Doleck, “Towards detecting and classifying malicious urls using deep learning,” J Wirel Mob Netw Ubiquitous Comput Dependable Appl, vol. 11, no. 4, pp. 31–48, Dec. 2020, doi: 10.22667/JOWUA.2020.12.31.031.

[10] Y. Liang and X. Yan, “Using deep learning to detect malicious URLs,” in Proceedings - IEEE International Conference on Energy Internet, ICEI 2019, Institute of Electrical and Electronics Engineers Inc., May 2019, pp. 487–492. doi: 10.1109/ICEI.2019.00092.

[11] Y. Liang and X. Yan, “Using deep learning to detect malicious URLs,” in Proceedings - IEEE International Conference on Energy Internet, ICEI 2019, Institute of Electrical and Electronics Engineers Inc., May 2019, pp. 487–492. doi: 10.1109/ICEI.2019.00092.

[12] T. Rasymas and L. Dovydaitis, “Detection of phishing URLs by using deep learning approach and multiple features

combinations," Baltic Journal of Modern Computing, vol. 8, no. 3, pp. 471–483, 2020, doi: 10.22364/BJMC.2020.8.3.06.

[13] I. Thakur, K. Panda, and S. Kumar, "Deep learning methods for malicious URL detection using embedding techniques as Logistic Regression with Lasso penalty and Random Forest," in PDGC 2022 - 2022 7th International Conference on Parallel, Distributed and Grid Computing, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 181–186. doi: 10.1109/PDGC56933.2022.10053199.

[14] U. R. Seshasayee, Arun Manoharan Hemprasad Patil Sujatha Rajkumar, vol. 1. 2019.

[15] A. Assefa and R. Katarya, "Intelligent Phishing Website Detection Using Deep Learning," in 8th International Conference on Advanced Computing and Communication Systems, ICACCS 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 1741–1745. doi: 10.1109/ICACCS54159.2022.9785003.

[16] M. Aljabri and S. Mirza, "Phishing Attacks Detection using Machine Learning and Deep Learning Models," in Proceedings - 2022 7th International Conference on Data Science and Machine Learning Applications, CDMA 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 175–180. doi: 10.1109/CDMA54072.2022.00034.

[17] P. Rastogi, E. Singh, V. Malik, A. Gupta, and S. Vijh, "Detection of Malicious Cyber Fraud using Machine Learning Techniques," in Proceedings of the Confluence 2022 - 12th International Conference on Cloud Computing, Data Science and Engineering, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 520–524. doi: 10.1109/Confluence52989.2022.9734181.

[18] Z. Peng, Y. He, Z. Sun, J. Ni, B. Niu, and X. Deng, "Crafting Text Adversarial Examples to Attack the Deep-Learning-based Malicious URL Detection," in IEEE International Conference on Communications, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 3118–3123. doi: 10.1109/ICC45855.2022.9838536.

[19] D. K. Karnase, "International Journal on Recent and Innovation Trends in Computing and Communication A Review on Malicious URL Detection using Machine Learning Systems," 2018, [Online]. Available: http://www.ijritcc.org

[20] H. Le, Q. Pham, D. Sahoo, and S. C. H. Hoi, "URLNet: Learning a URL Representation with Deep Learning for Malicious URL Detection," Feb. 2018, [Online]. Available: http://arxiv.org/abs/1802.03162

[21] S. Srinivasan, R. Vinayakumar, A. Arunachalam, M. Alazab, and K. P. Soman, "DURLD: Malicious URL detection using deep learning-based character level representations," in Malware Analysis Using Artificial Intelligence and Deep Learning, Springer International Publishing, 2020, pp. 535–554. doi: 10.1007/978-3-030-62582-5_21.

[22] V. Vundavalli, F. Barsha, M. Masum, H. Shahriar, and H. Haddad, "Malicious URL Detection Using Supervised Machine Learning Techniques," in ACM International Conference Proceeding Series, Association for Computing Machinery, Nov. 2020. doi: 10.1145/3433174.3433592.

[23] C. Do Xuan, H. Dinh Nguyen, and T. Victor Nikolaevich, "Malicious URL Detection based on Machine Learning," 2020. [Online]. Available: www.ijacsa.thesai.org

[24] S. Haque, Z. Eberhart, A. Bansal, and C. McMillan, "Semantic Similarity Metrics for Evaluating Source Code Summarization," in IEEE International Conference on Program Comprehension, IEEE Computer Society, 2022, pp. 36–47. doi: 10.1145/nnnnnnn.nnnnnnn.

[25] Miss. M. Pohane and Dr. A. A. Bardekar, "Review Paper on Detection of Malicious URLs Using Machine Learning Techniques," Int J Res Appl Sci Eng Technol, vol. 10, no. 3, pp. 2313–2314, Mar. 2022, doi: 10.22214/ijraset.2022.41065.

[26] M. Aljabri et al., "Detecting Malicious URLs Using Machine Learning Techniques: Review and Research Directions," IEEE Access, vol. 10, pp. 121395–121417, 2022, doi: 10.1109/ACCESS.2022.3222307.

[27] M. Alazab and S. Fellow, "Malicious URL Detection using Deep Learning."

[28] Shantanu, B. Janet, and R. Joshua Arul Kumar, "Malicious URL Detection: A Comparative Study," in Proceedings - International Conference on Artificial Intelligence and Smart Systems, ICAIS 2021, Institute of Electrical and Electronics Engineers Inc., Mar. 2021, pp. 1147–1151. doi: 10.1109/ICAIS50930.2021.9396014.