Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 2 : 2024
ISSN :**1906-9685**

MULTI-MODAL FEATURES REPRESENTATION-BASED CONVOLUTIONAL NEURAL NETWORK MODEL FOR MALICIOUS WEBSITE DETECTION

K. JAYARAJAN [1], B. NAVYA SRI[2], CH. SATHWIKA[3], B.S. SARVANI[4]
PROFESSOR 1, UG SCHOLAR 2,3,4
DEPARTMENT OF IT, MALLAREDDY COLLEGE OF ENGINEERING FOR WOMEN

ABSTRACT

Web applications have proliferated across various business sectors, serving as essential tools for billions of users in their daily lives activities. However, many of these applications are malicious which is a major threat to Internet users as they can steal sensitive information, install malware, and propagate spam. Detecting malicious websites by analyzing web content is ineffective due to the complexity of extraction of the representative features, the huge data volume, the evolving nature of the malicious patterns, the stealthy nature of the attacks, and the limitations of traditional classifiers. Uniform Resource Locators (URL) features are static and can often provide immediate insights about the website without the need to load its content. However, existing solutions for detecting malicious web applications through web content analysis often struggle due to complex feature extraction, massive data volumes, evolving attack patterns, and limitations of traditional classifiers. Leveraging solely lexical URL features proves insufficient, potentially leading to inaccurate classifications. This study proposes a multimodal representation approach that fuses textual and image-based features to enhance the performance of the malicious website detection. Textual features facilitate the deep learning model's ability to understand and represent detailed semantic information related to attack patterns, while image features are effective in recognizing more general malicious patterns. In doing so, patterns that are hidden in textual format may be recognizable in image format. Two Convolutional Neural Network (CNN) models were constructed to extract the hidden features from both textual and imagerepresented features. The output layers of both models were combined and used as input for an artificial neural network classifier for decision-making. Results show the effectiveness of the proposed model when compared to other models. The overall performance in terms of Matthews Correlation Coefficient (MCC) was improved by 4.3% while the false positive rate was reduced by 1.5%.

INTRODUCTION

According to the Siteefy website [1], there are over 1.11 billion websites in the World, and this number has been growing exponentially in recent years. Every day, T 252 thousand new websites are created (REF Please). As of May 9, 2023, it is estimated that the number of web pages is more than 50 billion pages. Although most of the websites are created for good purposes, many of these websites are malicious websites [2]. Malicious websites are designed to harm users in some way, such as by stealing their personal information or installing malware on their computers. They can be used tospread malware, phishing, spread spam, or conduct denial of service attacks [3]. According to Google's in-depth research, there are an estimated 12.8 million malicious websites on the internet [4]. Furthermore, as stated by authors in [5], there are 18.5 million websites hosting malicious code. This number is constantly changing, as new malicious websites are created and old ones are taken down. Malicious website detection has been the subject of much research and many solutions were suggested blacklist is the most common solution used by many organizations [24]. However, it is slow to update, as malicious actors can easily bypass blacklists by creating new websites or simply changing the URLs of their websites. This makes it difficult for blacklist-based systems to keep up with the ever-changing landscape of malicious websites. To address the limitations of blacklisting, many researchers have employed machine learning techniques to detect malicious websites. These techniques extract features from web content scripts HTTP/s response URLs domain names network traffic data and digital certificates. Many machine learning algorithms were used such as support vector machines, decision trees, logistic regression, and random forests to classify websites as malicious or benign. The effectiveness of machine learning methods depends on the choice

https://doi.org/10.36893/JNAO.2024.V15I12.051

of features. However, extracting effective features is challenging due to the constant changing of malicious code, the use of obfuscation techniques by attackers, the huge volume of data that needs to be analyzed, and the complexity of the attack today. Unfortunately, traditional machine learning is ineffective in extracting useful patterns for classification from huge and complex datasets. However, effective feature engineering is required to improve detection performance. Deep learning models are effective in extracting representative features from huge and complex datasets. They can automatically extract effective features without the need for incentive manual feature engineering, as it can automatically learn features from webpage text data. Convolutional Neural Networks (CNN) [22], Recurrent Neural Networks (RNN) , and attention mechanisms were commonly reported methods for malicious malware detection. Many deep learning models are constructed based on features extracted from the website's content. However, acquiring large and diverse datasets from website content for training deep learning models is challenging due to the dynamicity of the web content, the use of anti-scraping mechanisms to detect and block automated scrapers, and the evolving nature of online threats. Some websites require user sessions and authentication to access content. Scraping such websites may involve simulating user interactions, including logging in. Websites frequently change their structure and layout, necessitating ongoing maintenance and updates to scraping scripts to ensure they continue to work correctly. Moreover, extracting webpage representative features from the web content may be inefficient for limited resources devices such as IoT devices. Although content-based features can be used for detecting many types of threats, relying on web content features is neither effective nor efficient for detecting advanced malicious websites. The URL-based features seem to be a good alternative to the web content features. Many researchers compare the performance of the models constructed using both features and, on all occasions, URL-based features always win. However, most of the existing studies rely solely on the lexical features extracted from URLs. Lexical features have limited semantics information which causes the construction of sparse feature vectors. Some studies combine URL features with digital certificates to improve the detection performance. Malicious websites often lack valid certificates or use self-signed certificates, making certificate analysis a useful indicator of trustworthiness. Analyzing digital certificates can reveal whether a website is employing encryption, which is a common practice among reputable sites. However, not all websites use digital certificates, and some may employ self-signed certificates or certificates issued by less reputable Certificate Authorities (CAs). Extracting relevant and meaningful features from certificates for machine learning models can be complex, and the selection of the right features is crucial for effective detection. In addition, digital certificates can be misconfigured, expired, and frequently change leading to high false alarms. To sum up, existing solutions for detecting malicious web applications through web content analysis often struggle due to complex feature extraction, massive data volumes, evolving attack patterns, and limitations of traditional classifiers. Relying solely on lexical URL features proves insufficient, potentially leading to inaccurate classifications. To address these challenges, this study proposes a novel multimodal representation approach that integrates textual and image-based features to enhance malicious website detection. This approach leverages the strengths of both modalities: textual features capture detailed semantic information related to attack patterns, and image features recognize broader malicious visual cues. Hidden patterns within textual content may become discernible through image analysis. The proposed approach employs two Convolutional Neural Networks (CNNs): one for textual features and another for image features. Their outputs are then combined and fed into an artificial neural network classifier for improved decision-making. Our results demonstrate the superiority of the proposed model compared to existing approaches. We achieve a 4.3% increase in Matthews Correlation Coefficient (MCC) and a 1.5% reduction in the false-positive rate, showcasing the effectiveness of our multimodal approach in accurately identifying malicious web applications.

This study made the following contributions:

1. Integrating DNS-derived features with URL-based features enhances the comprehensiveness of malicious website detection. This synergy offers valuable contextual information regarding domain behavior and infrastructure, thereby fortifying the evaluation of website authenticity and security contributing to a more robust and nuanced approach to identifying malicious websites.
2. The study introduces a multimodal representation approach that utilizes both textual and image-based features to represent a

https://doi.org/10.36893/JNAO.2024.V15I12.051

comprehensive feature set. Textual features facilitate the deep learning model's ability to understand and represent detailed semantic information related to attack patterns, while image features are effective in recognizing more general malicious patterns.

3. Design and develop two Convolutional Neural Network (CNN) models to extract hidden features from the textual and image representations.

4. An additional, deep learning classifier was constructed to learn the relationships among the hidden features extracted by the CNN models. This approach advances the field by applying deep learning techniques to combine and leverage both textual and visual information for more effective malicious website detection.

LITERATURE REVIEW

1. J. McGahagan, D. Bhansali, C. Pinto-Coelho, and M. Cukier, ''Discovering features for detecting malicious websites: An empirical study,'' Comput. Secur., vol. 109, Oct. 2021, Art. no. 102374, doi: 10.1016/j.cose.2021.102374.

The paper focuses on identifying effective features for detecting malicious websites. With the rise of cybercrime, malicious websites pose significant threats, including phishing attacks, malware distribution, and fraud. Detecting these sites is crucial for enhancing online security. The authors conduct an empirical study to explore various features that can be used to distinguish malicious websites from legitimate ones. They analyze a broad set of features, including domain-based, content-based, and behavioral indicators, to determine which are most effective in identifying harmful websites. Through extensive data collection and experiments, the study identifies key features such as suspicious domain names, abnormal URL structures, and content characteristics that are indicative of malicious intent. The paper also evaluates different machine learning models for detecting these features and highlights the performance of classifiers when trained on the identified features. The results demonstrate that combining multiple feature types significantly improves detection accuracy, and the authors propose a comprehensive set of features that can be used to build more robust website classification systems. This research contributes to the field of cybersecurity by providing insights into the specific characteristics of malicious websites, which can be leveraged to develop more effective detection tools and improve online security measures. The findings emphasize the importance of feature discovery in combating cyber threats and suggest directions for future research in malicious website detection.

2. R. Patgiri, A. Biswas, and S. Nayak, ''DeepBF: Malicious URL detection using learned Bloom filter and evolutionary deep learning,'' Comput. Commun., vol. 200, pp. 30–41, Feb. 2023, doi: 10.1016/j.comcom.2022.12.027.

The paper introduces a novel method, *DeepBF*, for detecting malicious URLs. Malicious URLs are a critical threat in cybersecurity, often used for phishing, malware distribution, or other forms of cyberattacks. Traditional methods for URL detection often struggle with the high dimensionality of data and the evolving nature of cyber threats. To address these challenges, the authors propose combining a learned Bloom filter with evolutionary deep learning techniques. The learned Bloom filter is used to efficiently store and retrieve information about URL characteristics while maintaining low memory usage, thus improving the speed and scalability of the detection process. Evolutionary deep learning is employed to adaptively optimize the model, allowing it to learn and evolve as new malicious URLs emerge. The paper presents an innovative approach where the Bloom filter's parameters are learned through a deep learning framework, allowing the model to better capture the complex patterns of malicious URLs. The authors demonstrate the effectiveness of DeepBF through experiments that show its superior performance in detecting malicious URLs compared to traditional methods. The results highlight DeepBF's ability to balance detection accuracy with computational efficiency, making it a promising solution for real-time malicious URL detection in large-scale cybersecurity applications. This research contributes to the growing field of URL-based threat detection and emphasizes the potential of combining Bloom filters and evolutionary deep learning to enhance security systems.

3. M. Aljabri, H. S. Altamimi, S. A. Albelali, M. Al-Harbi, H. T. Alhuraib, N. K. Alotaibi, A. A. Alahmadi, F. Alhaidari, R. M. A. Mohammad, and K. Salah, ''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.

The paper provides a comprehensive review of machine learning techniques used for detecting malicious URLs. Malicious URLs are a major cyber threat, often used for phishing, malware distribution, and other forms of cyberattacks. The paper systematically analyzes the various machine learning methods employed in detecting these URLs, including supervised and unsupervised learning, deep learning, and hybrid models. It discusses the strengths and weaknesses of each technique, as well as the challenges involved in building robust malicious URL detection systems, such as handling large datasets, feature selection, and generalization across different attack vectors. The authors also highlight the importance of feature extraction,

exploring different URL-based features such as lexical, host-based, and content-based features that are essential for distinguishing malicious URLs from legitimate ones. Furthermore, the paper identifies key challenges in the field, including the need for real-time detection, scalability, and adaptability to new and evolving attack methods. The authors propose several future research directions, including the use of more advanced deep learning techniques, ensemble methods, and the incorporation of domain knowledge to improve detection accuracy and efficiency. By summarizing the current state of the art and offering insights into future trends, this paper provides valuable guidance for researchers and practitioners aiming to develop more effective and scalable solutions for malicious URL detection.

4. V. Devalla, S. S. Raghavan, S. Maste, J. D. Kotian, and D. D. Annapurna, ‘‘MURLi: A tool for detection of malicious URLs and injection attacks,’’ Proc. Comput. Sci., vol. 215, pp. 662–676, Jan. 2022, doi: 10.1016/j.procs.2022.12.068.

The paper presents MURLi, a tool designed to detect malicious URLs and injection attacks. Malicious URLs are a significant cybersecurity threat, often used in phishing, malware propagation, and other harmful activities. Injection attacks, such as SQL injection or script injection, are also commonly employed by attackers to exploit web applications and compromise system integrity. MURLi combines several detection techniques to identify malicious URLs and protect against injection attacks by analyzing the structure and content of URLs, as well as patterns of behavior that are indicative of harmful intent. The tool employs machine learning algorithms and heuristic methods to evaluate various characteristics of URLs, such as domain names, URL length, and the presence of suspicious keywords or patterns commonly found in malicious sites. In addition to detecting malicious URLs, MURLi also incorporates mechanisms to identify potential injection attacks, enhancing its ability to safeguard web applications. The authors demonstrate the tool's effectiveness through a series of experiments, showing that MURLi is capable of accurately detecting both malicious URLs and injection attack attempts with a high degree of reliability. The paper highlights the importance of such tools in defending against evolving cyber threats, particularly in the context of web security. MURLi represents a practical approach to improving cybersecurity by integrating detection capabilities for multiple types of attacks, offering a valuable solution for real-time threat mitigation.

5. H. Wang, Z. Tang, H. Li, J. Zhang, and C. Cai, ‘‘DDOFM: Dynamic malicious domain detection method based on feature mining,’’ Comput. Secur., vol. 130, Jul. 2023, Art. no. 103260, doi: 10.1016/j.cose.2023.103260.

The paper introduces a dynamic approach to detecting malicious domains using a feature mining technique. Malicious domains are commonly used in cyberattacks, such as phishing, malware distribution, and botnet activities, posing significant threats to cybersecurity. The paper presents the *DDOFM* (Dynamic Domain Detection Based on Feature Mining) method, which focuses on dynamically detecting malicious domains by extracting and analyzing a wide range of features that could indicate malicious intent. These features include domain characteristics, registration details, traffic patterns, and DNS resolution behaviors, all of which are indicative of malicious activity. The DDOFM approach adapts to the evolving nature of cyber threats by continuously updating its detection criteria, ensuring that it can identify new and emerging attack techniques. The authors employ machine learning techniques to mine and rank features from large-scale domain datasets, identifying the most relevant and impactful indicators of malicious activity. Experimental results demonstrate that DDOFM is effective in accurately detecting malicious domains while maintaining low false positive rates. The paper highlights the advantages of feature mining in domain detection, offering a more robust and scalable solution compared to traditional methods. By combining dynamic feature extraction with machine learning, the DDOFM method provides a powerful tool for real-time detection and prevention of domain-based cyber threats. This research contributes to the field of domain-based attack detection and emphasizes the importance of adaptive, feature-driven techniques in enhancing cybersecurity defenses.

EXISTING SYSTEM

There are three main approaches that have been suggested by researchers for malicious URL classification: blacklist, content-based, and URL-based [11], [32]. Many techniques were proposed to construct the detection classifiers such as the use of heuristic rules based on professional experience or the use of machine learning techniques. However, effective malicious URL detection is still an open issue. The performance of the recent malicious website detection solutions is influenced by the extracted features and the machine learning algorithms used for constructing the detection classifier. Authors in [32] presented an in-depth literature review that covers various machine learning-based techniques for detecting malicious URLs, considering aspects such as limitations, detection technologies, feature types, and datasets. The type of extracted features combined with deep learning techniques are research trends of malicious website detection solutions. The professional experience heuristic rule was widely used for constructing a blacklist of malicious URLs

such as the Google safe web browsing tool [37]. However, the blacklist solutions are ineffective for malicious URL detection due to the constantly evolving threats causing the need for frequent identification of the evolved threat and frequently updating the database.

- Many researchers have used feature extraction techniques to extract the features from website content to detect malicious content Natural language processing has been commonly employed for representation. However, due to the evolving nature of attacker's techniques, malicious website content is complex and such patterns become dynamic and stealthy leading to poor detection accuracy. For example, in [38], the authors investigated how malicious websites employ various web spam techniques to evade detection. The aim is to provide an effective solution for detecting and combating malicious websites that utilize techniques like redirection spam, hidden Iframes spam, and contenthiding spam. Accordingly, the study focuses on capturing screenshots of webpages from a user's perspective and using a Convolutional Neural Network for classification. However, the solution is limited for detecting spam techniques. Moreover, the feature depends on screenshots of the loaded page might be dangerous and uncompleted due to the dynamic nature of the websites.
- In [27], the authors collected features from the HTTP/s responses and applied various feature transformation and selection techniques for classification. However, these features are dynamic, subject to obfuscation using encoding and encryption mechanisms, which can render the detection classifier ineffective. Although machine learning algorithms were widely used for constructing the detection classifier, many researchers focused on deep learning techniques. Deep learning can accurately determine the similar patterns learned during the training resulting in effective classification. However, the web content is very dynamic and may be encrypted or encoded to hide the malicious patterns, posing a challenge in extracting effective features for classification.
- The URL features which less dynamic are promising for the accurate detection of malicious domains. This is because malicious domains are generated algorithmically while benign domains are created by humans. Thus, malicious URLs may contain more prominent features compared to the features extracted from the content which can be obfuscated, or encrypted to mislead the learning process. Authors in [38] focused on detecting the malicious URLs that are generated algorithmically. They hypothesize that attackers or malicious bots are used to generate the malicious URLs automatically. Accordingly, those URLs may contain patterns that are different from those generated by humans. Similarly, authors in [39] and [40] proposed solutions for detecting URLs that are generated using Domain Generation Algorithms (DGAs). Authors in [41] proposed a malicious website detection technique based on lexical and host-based features extracted from URLs. Results showed that URL features are more accurate compared to the other types of features.
- Authors in [26] proposed an adaptive segmentation mechanism to solve the maximum sequence length (MSL) limitation in deep learning. Webpage text, digital certificate, and Uniform Resource Locator (URL) were used as the source of the extracted features and used to construct the detection model using the Multi-Head Self-Attention and multi-channel text convolution (MCTC) network. However, relying on dynamic content features is challenging and can lead to degrade the classification performance. The study in [42] presented an approach to learning the uncertainties by employing deep Bayesian neural networks (DBNNs) to model the stochastic system dynamics. Authors in [43] presented a feature extraction algorithm called URL embedding based unsupervised learning technique called Huffman coding to reduce the dimensionality of the features vector. Although the algorithm shows better detection performance compared to the existing

https://doi.org/10.36893/JNAO.2024.V15I12.051

feature extraction mechanisms, the algorithm has been evaluated using a dataset with a strong assumption about the length and distribution of the characters of the malicious URLs samples.

- In [34], the authors proposed an anomaly detection model for detecting malicious domains. They utilized Hidden Markov Model (HMM) with a probabilistic model was used to construct the normal profile of the normal domain. In the online operation, if the domain is suspicious Jensen–Shannon divergence is calculated between the suspicious domain and a subset of the benign domains, and if the JS divergence exceeds a specific threshold the malicious domain is detected. Authors in [31] proposed a detection model called ‘‘deepBF’’ which combines Bloom Filters and Deep Learning techniques, aiming to improve accuracy and efficiency in identifying potentially harmful web addresses. The evolutionary convolutional neural network was used to construct the detection classifier. Authors in [33] compare the performance of several deep learning and traditional machine learning techniques to detect malicious URLs. The BiLSTM classifier was reported as the most performed classifier among studied classifiers. Authors in [21] used a combination of different feature transformations to reduce the data volume to improve the learning process. Various linear and non-linear space transformation methods were used in the solution. Although feature transformation plays a significant role in improving the classifiers constructed using traditional machine learning techniques, the total number of features extracted is 62 features does not seem very challenging if deep learning techniques were used for the classification.
- Authors in [44] presented a solution for malicious URL detection using two-stage ensemble learning to address the growing concern of web-based attacks. The study leverages cyber-threat intelligence features from sources like Google web search and Whois websites to enhance detection accuracy. The two-stage ensemble approach, combining Random Forest and Multi-Layer Perceptron algorithms, results in an improvement in accuracy and a reduction in false positives when compared to traditional URL-based models. However, the study does not thoroughly examine the potential limitations of relying on external cyber threat intelligence sources, which may pose challenges in terms of comprehensiveness and timeliness, warranting further investigation.

Disadvantages

In an Existing system, solutions for detecting malicious web applications through web content analysis often struggle due to complex feature extraction, massive data volumes, evolving attack patterns, and limitations of traditionalclassifiers. Relying solely on lexical URL features proves insufficient, potentially leading to inaccurate classifications.

PROPOSED SYSTEM

The system proposes a novel multimodal representation approach that integrates textual and image-based features to enhance malicious website detection. This approach leverages the strengths of both modalities: textual features capture detailed semantic information related to attack patterns, and image features recognize broader malicious visual cues. Hidden patterns within textual content may become discernible through image analysis.

The proposed approach employs two Convolutional Neural Networks (CNNs): one for textual features and another for image features. Their outputs are then combined and fed into an artificial neural network classifier for improved decision-making. Our results demonstrate the superiority of the proposed model compared to existing approaches. We achieve a 4.3% increase in Matthews Correlation Coefficient (MCC) and a 1.5% reduction in the false-positive rate, showcasing the effectiveness of our multimodal approach in accurately identifying malicious web applications.

Advantages

1. Integrating DNS-derived features with URL-based features enhances the comprehensiveness of malicious website detection. This synergy offers valuable contextual information regarding domain behavior and infrastructure, thereby fortifying the evaluation of website authenticity and security contributing to a more

robust and nuanced approach to identifying malicious websites.

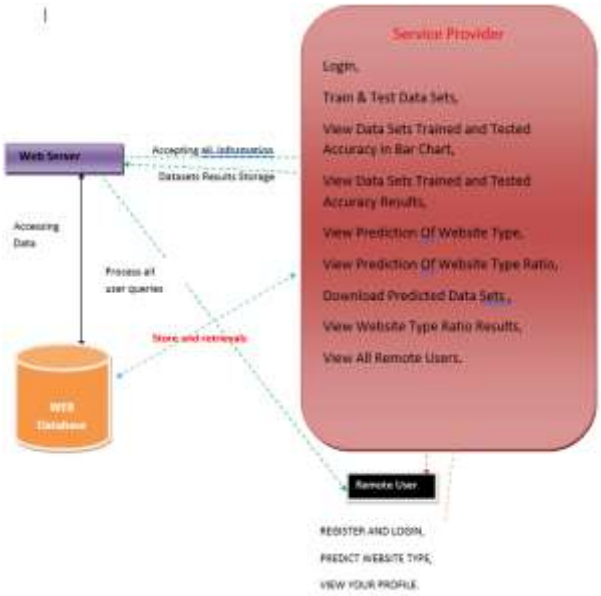
https://doi.org/10.36893/JNAO.2024.V15I12.051

2. The study introduces a multimodal representation approach that utilizes both textual and image-based features to represent a comprehensive feature set. Textual features facilitate the deep learning model's ability to understand and represent detailed semantic information related to attack patterns, while image features are effective in recognizing more general malicious patterns.

3. Design and develop two Convolutional Neural Network (CNN) models to extract hidden features from the textual and image representations.

4. An additional, deep learning classifier was constructed to learn the relationships among the hidden features extracted by the CNN models. This approach advances the field by applying deep learning techniques to combine and leverage both textual and visual information for more effective malicious website detection.

IMPLEMENTATION
SYSTEM ARCHITECTURE



MODULES
SERVICE PROVIDER

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Browse and Train & Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Student Teacher Network Attack Status, View Student Teacher Network Attack Status Ratio, Download Trained Data Sets, View Student Teacher Network Attack Status Ratio Results, View All Remote Users.

VIEW AND AUTHORIZE USERS

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

REMOTE USER

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, Find Student Teacher Network Attack Status, VIEW YOUR PROFILE.

RESULT

https://doi.org/10.36893/JNAO.2024.V15I12.051

CONCLUSION

In this study, a malicious website detection model called HF-CNN was designed and developed. The model integrates URL features with DNS features to enhance the comprehensiveness of identifying malicious websites. A multimodal representation approach that encompasses both textual and image-based characteristics has been proposed to depict the combined feature set. Textual attributes enable the deep learning model to grasp and depict complex semantic details associated with attack patterns, while image attributes surpass at recognizing broader malicious patterns. Two Convolutional Neural Network (CNN) models were constructed to extract hidden features from the textual and image representations. CNNs are capable of simultaneously capturing both local and global features. The results indicate that the proposed model outperforms the other related models. The overall performance in terms of F-measure and MCC has been improved by 0.4%, and 0.6%, respectively, compared with the baseline model txt CNN. The False Positive Rate (FPR) and False Negative Rate (FNR) were reduced by 1.6% and 1.4%, respectively. While the proposed models achieved a high detection performance of 98.88% in terms of F-measure, there are still considerable amounts of errors presented in the detection performance as measured by the MMC score of 96.66%. The errors mostly resulted from the unrepresented features in URLs and DNS information. Therefore, relying solely on URLs, DNS information or static features is not a wise approach to malicious website detection, as some benign domains that suffer from security vulnerabilities may become malicious due to injection attacks. Therefore, it is important to combine the URL-based features with other features such as content features. However, content features are complex due to their high dynamicity and usability by attackers to evade detection. As a result, further research is needed to propose effective and efficient mechanisms for acquiring web content.Furthermore, employing an adaptive ensemble of classifiers designed to accommodate the dynamic nature of evolving threats could enhance detection performance. Each classifier within the ensemble is constructed based on a distinct set of features, providing versatility and robustness in addressing diverse threat scenarios.

REFERENCES

[1] NJ. (2023). How Many Websites are There in the World? Accessed: Sep. 10, 2023. [Online]. Available: https://siteefy.com/how-manywebsites-are-there/

[2] M. Liu, B. Zhang, W. Chen, and X. Zhang, ‘‘A survey of exploitation and detection methods of XSS vulnerabilities,’’ IEEE Access, vol. 7, pp. 182004–182016, 2019.

[3] J. Jang-Jaccard and S. Nepal, ‘‘A survey of emerging threats in cybersecurity,’’ J. Comput. Syst. Sci., vol. 80, no. 5, pp. 973–993, Aug. 2014, doi: 10.1016/j.jcss.2014.02.005.

[4] N. Provos, D. McNamee, P. Mavrommatis, K. Wang, and N. Modadugu, ‘‘The ghost in the browser: Analysis of web-based malware,’’ HotBots, vol. 7, p. 4, Apr. 2007.

[5] K. Townsend, ‘‘18.5 Million websites infected with malware at any time,’’ Wired Bus. Media, SecurityWeek, Boston, MA, USA, Tech. Rep. Q4 2017, 2022. Accessed: Feb. 1, 2022. [Online]. Available: https://www.securityweek.com/185-million-websites-infected-malwareany-time

[6] A. S. Raja, R. Vinodini, and A. Kavitha, ‘‘Lexical features based malicious URL detection using machine learning techniques,’’ Mater. Today, vol. 47, pp. 163–166, Jan. 2021, doi: 10.1016/j.matpr.2021.04.041.

[7] A. Subasi, M. Balfaqih, Z. Balfagih, and K. Alfawwaz, ‘‘A comparative evaluation of ensemble classifiers for malicious webpage detection,’’ Proc. Comput. Sci., vol. 194, pp. 272–279, Jan. 2021, doi: 10.1016/j.procs.2021.10.082.

[8] S. R. Zahra, M. A. Chishti, A. I. Baba, and F. Wu, ‘‘Detecting COVID19 chaos driven phishing/malicious URL attacks by a fuzzy logic and data mining based intelligence system,’’ Egyptian Informat. J., vol. 23, no. 2, pp. 197–214, Jul. 2022, doi: 10.1016/j.eij.2021.12.003.

[9] B. B. Gupta, K. Yadav, I. Razzak, K. Psannis, A. Castiglione, and X. Chang, ‘‘A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment,’’ Comput. Commun., vol. 175, pp. 47–57, Jul. 2021, doi: 10.1016/j.comcom.2021.04.023.

[10] R. Wazirali, R. Ahmad, and A. A.-K. Abu-Ein, ‘‘Sustaining accurate detection of phishing URLs using SDN and feature selection approaches,’’ Comput. Netw., vol. 201, Dec. 2021, Art. no. 108591, doi: 10.1016/j.comnet.2021.108591.

[11] D. K. Mondal, B. C. Singh, H. Hu, S. Biswas, Z. Alom, and M. A. Azim, ‘‘SeizeMaliciousURL: A novel learning approach to detect malicious URLs,’’ J. Inf. Secur. Appl., vol. 62, Nov. 2021, Art. no. 102967, doi: 10.1016/j.jisa.2021.102967.

[12] K. Haynes, H. Shirazi, and I. Ray, ‘‘Lightweight URL-based phishing detection using natural language processing transformers for mobile devices,’’ Proc. Comput. Sci., vol. 191, pp. 127–134, Jan. 2021, doi: 10.1016/j.procs.2021.07.040.