Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

SICAIET
2024

Research Article
Collection: Artificial Intelligence and Emerging Technologies

Enhancing Phishing Detection, Leveraging Deep Learning Techniques

Azmat Ullah[1], Rizwan Ali Shah[1], Syed Ali Nawaz[1*], Nazeer Ahmad[1], and Mubasher H. Malik[2]

[1]Department of Information Technology, The Islamia University of Bahawalpur (IUB), Bahawalpur 63100, Pakistan.
[2]Department of Computer Science, Institute of Southern Punjab, Multan, Pakistan.
[*]Corresponding Author: Syed Ali Nawaz Email addresses: ali.nawaz@iub.edu.pk

Academic Editor: Salman Qadri **Published:** February 01, 2024

Abstract: The widespread usage of internet-connected gadgets has led to a transformation in how individuals engage with technology, facilitating easy participation in various online activities such as social media, banking, and shopping. However, this proliferation has also provided an opportunity for fraudsters to exploit the Internet's anonymity through elaborate phishing schemes. These schemes aim to deceive users into disclosing personal information, including passwords and bank account details, often employing social engineering techniques. Consequently, the development of sophisticated phishing detection systems has become imperative to safeguard users' financial assets and digital identities. Many of these systems leverage state-of-the-art technology, with a significant reliance on machine learning methods for accurate and rapid detection of phishing attempts. Among these methods, deep learning algorithms have gained prominence due to their ability to efficiently process and analyze vast amounts of data. This paper presents a unique phishing detection system grounded in deep learning principles, employing a diverse array of techniques such as CNNs, attention networks, neural architects, and recurrent neural networks. The system's efficacy in real-time phishing attack detection is demonstrated through its focus on swift categorization of web pages based on URLs. Evaluation of the proposed method using a large dataset comprising nearly five million tagged URLs underscores its effectiveness.

Keywords: Cyber Security; Phishing Detection; Phishing; Anti Phishing; Security; CNNs; RNN.

1. Introduction:

From online banking to e-commerce, the internet powers a myriad of essential activities in today's digital era. In light of these everyday benefits, however, the internet also poses security vulnerabilities that hackers exploit — and one of the most common strategies they've embraced is phishing. Phishing is an illegal tactic employed by individuals that involves attempting to acquire personal information, often under the guise of legitimate emails or services.

Phishing assaults can be highly damaging to both individuals and organizations. They frequently come in the form of fraudulent emails or deceitful websites that pose as those of respected organizations in an attempt to hoodwink people into divulging passwords, bank account numbers and other personal information. However, despite the extension of various protections, individuals are defenseless to phishing attacks for a number of reasons. One of the main reasons is that real Uniform Resource Locators (URLs), the unique web address used to access a website, are often difficult to tell apart from their fakes — and not just by users. Cybercriminals can also take advantage of the former because they can associate harmful URLs so effectively with innocent URLs, and users are more quick to trust URLs that look innocent.

Phishing attacks are a reminder that online safety remains quite complex, even as technology and our reliance on it continues to evolve at a breakneck pace. They are also a reminder that individuals still need to be proactive in protecting their personal information. This requires using strong security practices and technology to protect personal information, recognizing common phishing tactics and knowing what actions to take if they suspect that they have encountered a phishing attempt. By being alert and taking the necessary precautionary steps, people can reduce their likelihood of falling victim to phishing attempts and help lay the foundation for a safer online experience overall.

As phishing tactics develop, they increasingly take advantage of end users, who are the weakest link in the security chain. Raising user awareness and creating reliable detection and prevention tools are therefore essential components of a two-layered security approach. Because they can adapt to changing attack strategies, machine learning-based solutions, especially those that make use of deep learning technologies, have become strong rivals among the many phishing detection options that are now available.

In order to demonstrate the effectiveness of deep learning algorithms for real-time phishing detection, this article focuses on recognizing phishing attempts that were previously unknown, including zero-day threats. It is noteworthy that the proposed system functions independently, ensuring quick detection and demonstrating its self-sufficiency. When compared to traditional approaches, experimental results show a considerable reduction in false positives and significant improvements in phishing detection accuracy.

In the paper, the researchers announced significant advance- ments made in cybersecurity related to phishing detec- tion. A major contribution of the study was the creation of a large, well managed dataset that is critical for training and testing phishing detection systems. The integrity of detection algorithms is ensured by the coverage of phishing scenarios and the even distribution across a variety of types and features in the dataset.

The study also demonstrates novel strategies for enhancing the efficiency of phishing detection systems. The system demonstrates for the first time

integration of advanced algorithms and techniques, reducing false positives and strengthening defenses overall. These developments reduce the vulnerability to phishing scams by increasing the con- fidence of detection systems.

Further, the paper illustrated the first use of URL analysis for fast detection using site domain features, which lets the system rapidly identify and classify any phishing attacks. Real-time data analysis is critical to proactively block phishing attempts and shield con- sumers and businesses from security breaches and theft of sensitive data.

The system's capacity to successfully identify ze-ro-day assaults is another noteworthy advancement in cybersecurity. The solution reduces the impact of new assaults on security systems by effectively recognizing phishing threats that were previously undetected, giving it an advantage over constantly changing attack techniques.

The paper's remaining sections describe experimental data, make conclusions, offer background information, explore related research, highlight the complexities of the suggested system, and suggest future study options.

Phishing detection has advanced significantly thanks to the paper's contributions, which include language-independent capabilities, fast analysis, zero-day detection, dataset generation, and detection improvement. Through their contributions, security protocols may be strengthened and resistance to constantly changing cyberthreats may be increased.

This study presents a unique phishing detection system that makes use of deep learning principles in response to the rising threat posed by phishing attempts in the digital age. To handle and assess large volumes of data effectively, the system includes state-of-the-art technologies such as recurrent neural networks (RNNs), neural architects, attention networks, and convolutional neural networks (CNNs). In contrast to traditional techniques, our technology concentrates on quickly classifying webpages using URL analysis, allowing for the real-time identification of any phishing attempts. We have made great progress in fighting cyber threats with our special phishing detection technology, which offers users better defense against ever-changing phishing techniques.

The main contributions of my study are:

- Real time detection of phishing attempts and prompt actions.
- Integration of multiple techniques.
- Robust phishing detection system capable of quickly identifying suspicious URLs
- Integration of multiple techniques.

2. Materials and Methods

2.1 Review of Related Research

The surge in phishing emails, including spam, spear phishing, and malware-infected messages, underscores the critical need for reliable, intelligent anti-spam filters. This survey paper offers a focused examination of AI and ML techniques for smart spam detection, aiming to support the development of effective defenses. Our analysis concentrates on four email structure components: (A) headers, providing routing details such as sender and recipient email and IP addresses; (B) SMTP envelope, with mail exchanger identifiers and source/destination domains/users; (C) initial SMTP data segment, including sender, recipient, date, and subject; and (D) subsequent SMTP data segment, encompassing email content and attachments. Through evaluating emerging intelligent methods, we identify, review, and summarize relevant literature. Furthermore, the paper uncovers valuable insights, discusses challenges, and outlines research gaps, serving as a guide for future investigations to enhance spam detection techniques [1].

Phishing attackers use email, text messages, and social media to trick users into revealing personal data, which they exploit for fraudulent activities. The rise of machine learning has led to various phishing detection methods, including the deep learning-based framework proposed in this paper. Implemented as a browser plug-in, it assesses real-time phishing risks and employs multiple strategies, including ML prediction, to enhance accuracy. Experimental results show the RNN-GRU model achieves 99.18% accuracy [2].

Cloud computing is a cost-effective innovation that reduces upfront hardware and computing expenses. Fog computing enhances cloud infrastructure by optimizing end-user response times, particularly for IoT applications. Despite these benefits, IoT devices are vulnerable to cyber attacks like DoS, DDoS, and botnets. To combat this, we propose a novel botnet detection method suitable for fog computing, leveraging software-defined networking (SDN). Our method, rigorously tested on recent datasets, achieves superior performance, accurately identifying 99.98% of sophisticated bot attacks with a processing time of 0.022 milliseconds [3].

Phishing, an online social engineering attack, involves extracting valuable data from unsuspecting users. Despite the effectiveness of current anti-phishing tools, hackers continuously devise new evasion tactics. Given the severe impact of phishing, there's a pressing need for innovative countermeasures. While artificial intelligence (AI) is crucial in combating phishing, its methods often struggle with high false alarms and understanding phishing techniques. This study introduces four meta-learner models based on the extra-tree classifier to tackle these issues. Trained on updated datasets, these models consistently achieve detection accuracy above 97% with a very low false-positive rate of under 0.028, outperforming existing machine learning models. Therefore, integrating meta-learners is recommended for developing effective phishing detection systems [4].

We conduct an extensive benchmarking analysis of phishing features across diverse datasets, proposing a taxonomy of features and the 'PhishBench' framework. PhishBench facilitates the systematic evaluation and comparison of existing features for phishing detection under standardized conditions. It represents a pioneering effort in this domain, enabling thorough evaluation and feature comparison. Using PhishBench, we assess methods from the literature to evaluate their robustness and scalability. Our investigation highlights how dataset characteristics impact classification performance, emphasizing the need to address the imbalanced nature of phishing attacks. Additionally, we find that new features and techniques are essential to enhancing detection system effectiveness against evolving attacks [5].

Crypto-ransomware, a predominant modern malware, heavily impacts industries like small businesses, healthcare, education, and government, often demanding significant ransoms. Existing static and dynamic analysis techniques struggle against evolving malware. Our study introduces an AI-powered hybrid approach for ransomware detection, integrating advanced methods with behavioral chain analysis and machine learning. Experimentation yielded a machine learning algorithm with 99.72% accuracy and a 0.003 false positive rate. Multi-level profiling enhances detection accuracy across various operating systems. We developed AIRaD (AI-based ransomware detection) to assist in result visualization and interpretation for researchers and defenders [6].

This research introduces an innovative approach to identifying malicious URLs by employing a simulated expert (SE) and a knowledge base system (KBS). The proposed methodology not only effectively detects known malicious URLs but also adapts countermeasures against newly generated ones.

Additionally, the study examines the significance of various lexical features in the decision-making process, utilizing factor analysis to reduce dependency on human experts. Furthermore, the performance of several state-of-the-art machine learning algorithms is evaluated using a large-scale real-world dataset of web applications. Experimental findings reveal that Naïve Bayes (NB) outperforms other algorithms, demonstrating an average execution time of 3 seconds. Moreover, it accurately classifies 107,586 URLs with a mere 0.2% error rate and a 99.8% accuracy rate [7].

Cyberattacks pose challenges in intrusion detection, risking data security. This review explores recent IDS developments, evasion tactics, and efforts to bolster network security. Researchers strive to enhance IDS accuracy, leveraging ML and DL techniques for intrusion detection. The paper discusses trends in ML and DL-based NIDS, addressing methodologies and challenges. It also proposes a research model to tackle methodological weaknesses and suggests using decision trees for anomaly detection. This research aims to guide the development of a robust detection framework [8].

In recent years, the Internet has witnessed significant growth, accompanied by an increase in cyberattack risks, particularly from malicious URLs. These URLs aim to extract unauthorized information, posing substantial threats to users and necessitating enhanced website security measures. This paper reviews machine learning-based techniques for detecting malicious URLs, discussing limitations, detection technologies, features, and datasets. Additionally, it highlights research gaps in detecting malicious Arabic websites and identifies challenges in URL detection, proposing solutions [9].

Machine learning and deep learning techniques are integral to financial operations, aiding in trading, mobile banking, and fraud detection. As financial crime shifts to cyberspace, termed "financial cybercrime," detection becomes complex. Current security measures face challenges from hackers and social engineers. Addressing transparency and fairness demands further complicates fraud detection. While traditional methods persist, newer graph-based and neural network models emerge. This survey aims to explore this evolving landscape, covering fraudulent methods, detection strategies, stakeholders, and emerging challenges [10].

Phishing remains a prevalent threat, with attackers continuously evolving techniques to deceive users and obtain sensitive information. Given the challenge for users to discern legitimate websites amidst these evolving tactics, a legitimate website detection system becomes crucial. To address this, we propose a hybrid deep learning model, HCNN-LSTM, combining convolutional neural networks (CNN) and long short-term memory (LSTM). Implemented on a benchmark dataset comprising 11,430 URLs, the model achieved high accuracy (95.19%), precision (95.00%), recall (95.00%), and F1-score (95.00%). Outperforming both CNN and LSTM models, our approach presents a competitive solution for legitimate website prediction tasks [11].

Network attacks exploit vulnerabilities to compromise security, posing risks like unauthorized access and data breaches. Common attacks include phishing and malware distribution, leading to financial losses. To combat these, organizations use intrusion detection systems (IDS) with advanced algorithms and machine learning. Our research focuses on timely and effective attack detection using the CPRF approach, achieving a remarkable accuracy of 99.9% [12].

Phishing emails represent a significant worldwide menace, resulting in considerable financial harm and exhibiting an upward trend in recent times. In response, we introduce THEMIS, an innovative phishing email detection model built upon an improved RCNN architecture. THEMIS conducts thorough scrutiny of email structures, utilizing multilevel vectors and attention mechanisms. Assessment on a real-world dataset demonstrates that THEMIS attains an accuracy of 99.848% and a false positive rate of 0.043%, outperforming existing approaches in the effectiveness of phishing detection [13].

Phishing, a form of cyberattack based on social engineering, seeks to deceive users into divulging credentials through deceptive websites. This research examines various machine learning and deep learning techniques for detecting phishing websites by analyzing their URLs. Unlike existing methods that often disregard login pages, we incorporate them into our analysis, shedding light on the significant issue of high false-positive rates. Our investigation involves the examination of datasets from different time periods, revealing a decline in model accuracy over time. Additionally, we perform a frequency analysis of phishing domains. To validate our findings, we introduce the PILU-90K dataset, which consists of 60,000 legitimate URLs (including login pages) and 30,000 phishing URLs. Finally, our logistic regression model, employing TF-IDF feature extraction, achieves an accuracy of 96.50% specifically for login URLs [14].

Phishing, a significant social engineering strategy, allows malicious actors to exploit end-users and infiltrate critical systems through emails containing embedded hyperlinks. Detecting and countering such attacks poses challenges due to their intricate and ever-changing nature. We present PhishLimiter, a novel approach that combines deep packet inspection (DPI) with software-defined networking (SDN) to detect and mitigate phishing attempts. Our DPI technique encompasses the classification of phishing signatures and real-time inspection. By harnessing SDN's programmability, we implement two modes—store and forward, and forward and inspect—to direct network traffic effectively. Using an artificial neural network model, PhishLimiter accurately identifies phishing attack signatures and adapts to evolving attack patterns. Moreover, PhishLimiter enhances network traffic management by leveraging SDN's comprehensive network view. Evaluation in a real-world testbed environment, utilizing datasets containing emails with embedded links, demonstrates PhishLimiter's effectiveness and efficiency in thwarting malicious activities [15].

As internet accessibility increases and our devices become more mobile, individuals are spending a greater amount of time on social networks. Unfortunately, cybercriminals are exploiting the popularity of these platforms by inundating them with spam, directing users to phishing sites or malware downloads, thereby posing significant safety risks and detracting from the overall user experience. Current solutions are encountering challenges in effectively identifying Twitter spam. This research endeavors to assess the performance of various mainstream machine learning algorithms using a substantial dataset, aiming to pinpoint those that offer accurate detection and consistent results. Furthermore, the study evaluates scalability to gauge real-time detection capabilities. Performance metrics encompass detection accuracy, true/false positive rates, and F-measure, while stability scrutinizes algorithm performance across randomly selected training samples of varying sizes. Additionally, scalability analysis delves into the impact of parallel computing on reducing training and testing time [16].

Ransomware has emerged as a significant cybersecurity threat, inflicting considerable financial, reputational, and operational damage on both individuals and organizations. This paper presents a thorough examination of ransomware's evolution, taxonomy, and current research landscape. We delve into its origins, classify various types, and scrutinize techniques for detection, prevention, mitigation, and prediction. Our comprehensive review, drawing from 150 references, highlights a predominant emphasis on ransomware detection (accounting for 72.8% of research efforts) compared to prediction methods. Machine learning features prominently in 70% of studies focused on detection. However, real-time security and zero-day ransomware identification continue to be difficult tasks, with adversarial machine learning and idea drift being two major obstacles. This survey serves as a valuable guide for researchers navigating the complex terrain of ransomware [17].

Phishing, an online attack to obtain sensitive information, challenges detection due to complex features. We propose a novel detection model using representation learning to extract multi-aspect features from URLs, HTML content, and DOM structure. Leveraging a hybrid deep learning network, our model achieves 99.05% accuracy and a 0.25% false positive rate, outperforming classic methods [18].

Heavy dependence on operational systems has led to numerous security issues. Detecting anomalies is crucial for safety, facilitated by anomaly diagnostic methods. However, cybersecurity faces challenges in data system construction due to the exploitation of valuable data, impacting data privacy. Additionally, attack incidents exacerbate these concerns [19].

Table 1. Comparative analysis of literature

Year /Ref	Applied Methods	Metrics (Acc.)	Pros	Cons
[20]	VAE and DNN	0.974	Not reliant on external services.	Long response time

[21]	LR, Light GBM, XGBoost, ADABoost, RF SVM, KNN, NB	0.965	Login URLs are employed independently, without relying on external services.	Without deep learning
[22]	NN, RF Averaged perception, SVM, LR and Boosted decision tree	0.977	Able to identify phishing emails.	Choosing a real-time system can be perplexing.
[23]	ANN, KNN, SVM, ANFIS	0.991	Utilize webpage content in addition to URL data.	Limited dataset for training and assessment.
[24]	RNN, CNN	0.978	fresh goal function	scant phishing data

2.2 Methodology

The best model selection would be a deep learning-based approach, specifically a recurrent neural network (RNN) coupled with a convolutional neural network (CNN) or transformer models.

By combining CNNs with RNNs or transformer models, the system can leverage the complementary strengths of both architectures to achieve robust phishing detection across different modalities and data types. This hybrid approach allows the system to effectively analyze and interpret diverse features, enhancing its ability to detect previously unknown phishing attacks, including zero-day threats.

Design a hybrid deep learning architecture that combines CNNs with either RNNs or transformer models. For the CNN-RNN combination,

- Utilize the CNN to extract features from images, screenshots, and logos embedded in phishing emails or websites.
- Feed the extracted features into an RNN, such as LSTM or GRU, to capture sequential patterns in text content and analyze the structure of emails, URLs, and website content.

Figure 1 prioritizes user awareness through frequent training and simulations, and mitigates phishing attempts. Figure 1 shows that blacklists and whitelists are used by software-based detection systems, such as list-based detection, to categorize URLs and IP addresses. Tools like Office 365's Advanced Threat Protection and Google Safe Browsing API are designed to quickly evaluate a URL's security before consumers engage with it.

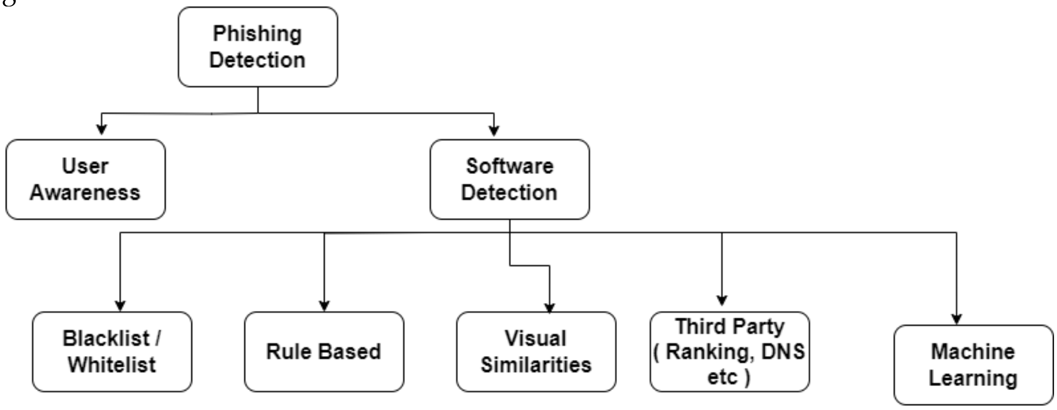


Figure 1. Model for detecting phishing attacks

The steps outlined above will be meticulously followed in the following sequence:

2.3 CNN Feature Extraction

- Initialize a pre-trained CNN model (e.g., VGG, ResNet) for image feature extraction.
- Pass each image embedded in phishing emails or websites through CNN.
- Extract the output features from one of the intermediate layers of the CNN.
- Flatten the extracted features to create a feature vector for each image.

2.4 RNN (LSTM/GRU) Input Preparation

- Tokenize the text content of phishing emails, URLs, and website content.
- Convert the tokens into numerical representations using techniques like word embedding (e.g., Word2Vec, GloVe).

- Pad or truncate the sequences to ensure uniform length across samples.

In the suggested model, we executed a data mining procedure according to the depicted flowchart in Figure 2. Firstly, we collected a relatively large dataset from diverse sources on the Internet. Subsequently, this dataset was divided into three segments to facilitate performance evaluation: 70% for training, 20% for validation, and 10% for testing. Equation 1 typically refers to the update equation of the hidden state in the network. RNNs are designed to process sequences of data, where each element in the sequence is processed sequentially, and the network maintains a hidden state that captures information about the past elements of the sequence. Equation 1 captures information about previous inputs by incorporating the current input along with the previous hidden state.

$$\mathrm{M_t = tanm(X_{mw}{}^{w_t} + \ X_{mm}{}^{m_{t-1}} + b_m\ )} \qquad (1)$$

$$\mathrm{y_t = softmax\ (\ X_{ym}{}^{m_t} + b_y} \qquad (2)$$

Equation 2 utilizes the softmax function, which calculates probabilities for individual classes using input scores. Following are the various components.

- $\mathrm{X_{mw}}$ is the input-to-hidden connections' weight matrix.
- $\mathrm{X_{mm}}$ is the weight matrix for links between concealed objects.
- $\mathrm{X_{ym}}$ is the connection's weight matrix from hidden to output.
- Bm and by are the bias vectors for the hidden and output layers, respectively.
- tanh is the hyperbolic tangent activation function.
- softmaxsoftmax is the softmax activation function.

We used the data mining procedure shown in Fig. 2 for our model. Employing this flowchart illustration, the five different deep learning algorithms—Artificial Neural Network, Convolutional Neural Network, Recurrent Neural Network, Bi-directional Recurrent Neural Network, and Attention Network were used to train the system by following this flowchart. The trained model may then be used in real-world scenarios to recognize phishing websites.

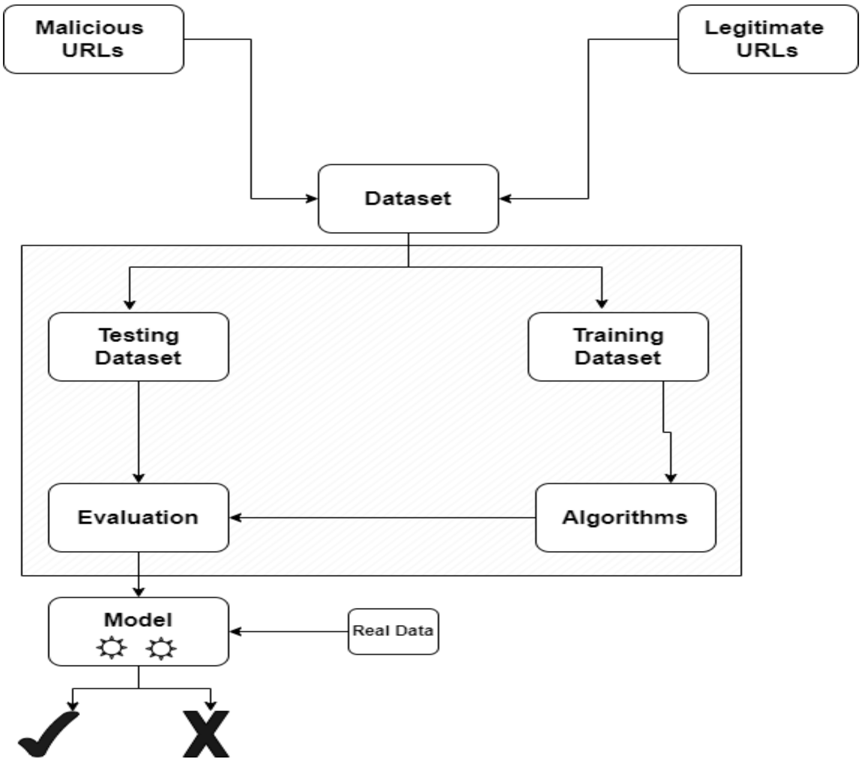


Figure 2. The flow diagram of Enhancing Phishing Detection: Leveraging Deep Learning Techniques

2.5 Dataset Collection

In this research, we assembled and employed a comprehensive dataset for conducting diverse experiments. Phishing URLs were collected from 'www.phishtank.com', while legitimate URLs were acquired from 'commoncrawl.org'. PhishTank, an open-source community platform, regularly updates and shares phishing URLs. These URLs are subject to a community voting process, which classifies them into categories such as valid phish, invalid Phish, or unknown, based on community consensus. The dataset consists of 2,320,893 phishing URLs sourced from PhishTank, specifically those shared until August 2018.

In building a machine learning application for detecting phishing attacks, it's essential to incorporate non-phishing URLs. To achieve this, we utilized the Common Crawl Corpus, which aggregates extensive data collected over seven years. This corpus comprises raw web page data, metadata, and textual content. Legitimate domains with substantial backlink counts typically receive higher rankings in search engine results, thereby bolstering the credibility of the dataset.

The legitimate URL dataset was compiled by extracting the top 100,000 domains from global PageRank rankings, resulting in a total of 2,881,948 legitimate URLs. The construction and assessment of phishing detection systems can be strengthened by this large dataset, which encompasses a wide range of genuine resources on the internet.

The monitoring system for phishing we created incorporates URLs from unquestionably legitimate domains. To prevent overfitting and ensure broad applicability, we included only a small number (10 or fewer) of URLs from these domains in the dataset. This approach allows the structure to accurately identify attempts at phishing across a wide range of legitimate domains and URL formats.

Table 2. Detail of the dataset

Dataset	Class	Small Dataset	Big Dataset
Train	Phish	162463	1624623
	Legitimate	201736	2017363
	Overall	364199	3641986
Validation	Phish	46649	466504
	Legitimate	57927	579271
	Overall	104576	1045774
Test	Phish	22978	229766
	Legitimate	28532	285314
	Overall	51510	515080
All	Phish	232090	2320893
	Legitimate	288195	2881948
	Overall	520285	5202841

Table 2 presents the biggest dataset in this subject, with 5.1 million URLs gathered (2.3 million phishing, 2.8 million legal). New data-driven iterative model upgrades improve detection. Sharing open-source materials encourages cooperation and transparency for ongoing development. Small_dataset (10% sample) and big_dataset (full dataset) are two dataset variants that account for hardware restrictions.

2.6 Vectorization:

Phishing attackers utilize a range of tactics, some of which pose notable hurdles for detection. These tactics frequently employ methods that complicate word-based analyses by incorporating subtle alterations that are challenging for individuals to detect. To conduct word-based analysis effectively, extensive training on a corpus containing a significant number of nonsensical words is necessary. Moreover, word-based analysis adds a degree of linguistic reliance. Hence, in this study, we chose to employ a method for vectorization that uses character-based embedding. Each character in the embedding is assigned a size of 50, resulting in the vectorized representation of the dataset as illustrated in Figure 3.

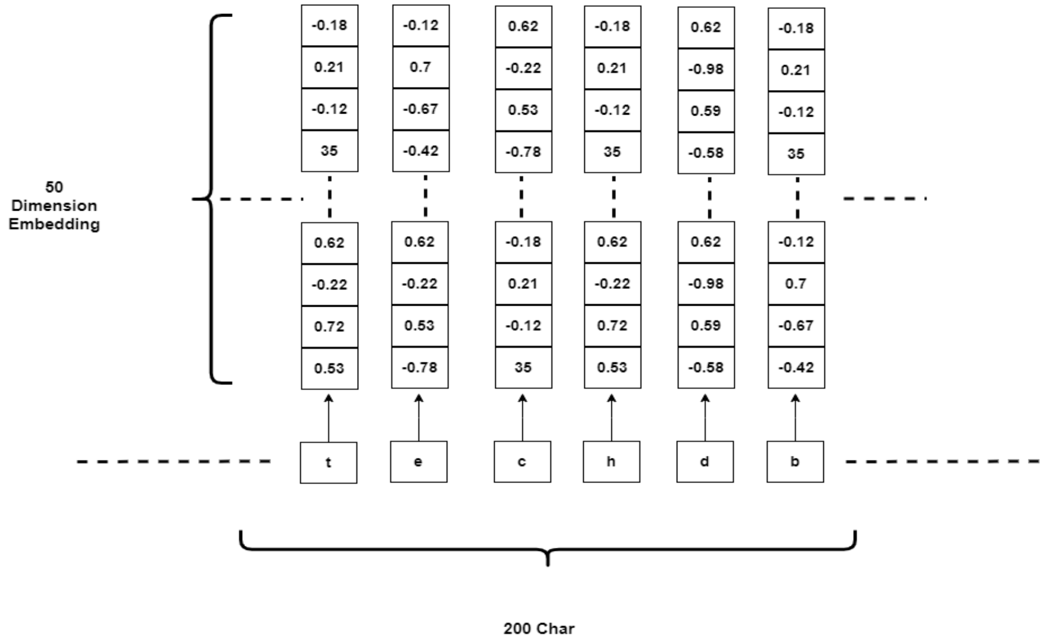


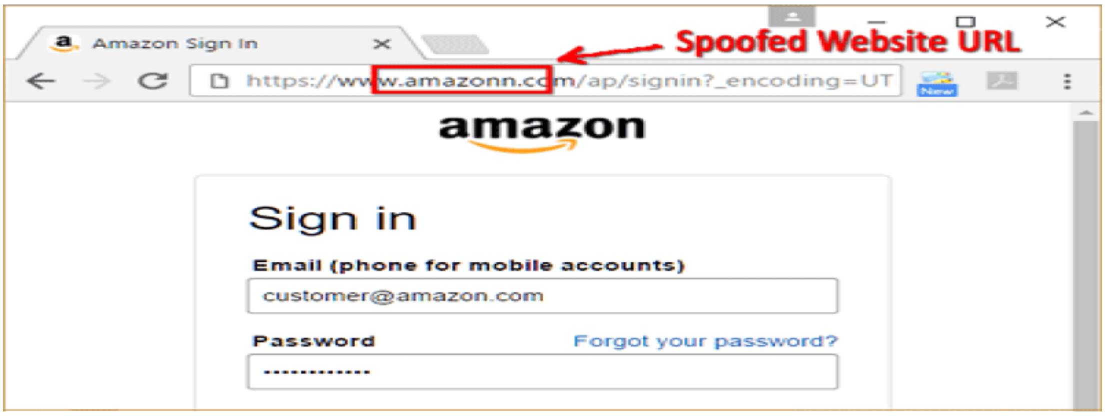
Figure 3. Vectorization procedure

In the course of the vectorization process, it is essential to confine the URL length to a predetermined number of characters. To determine the suitable character count, in Table 3 we analyze statistical metrics that consider the URLs' character counts. Subsequently, we compute the minimum, maximum, median, and standard deviation for each class separately. These statistical insights, derived from the comprehensive dataset, are presented in Table 3.

Table 3. Character frequency statistics for the dataset

Class	Count	Average	Min	Max	Std. Dev	Median
Phishing	2321934	86.9	12	6104	64.9	68
Legitimate	2881948	72.0	12	5984	45.3	63
All	5203882	78.6	12	6104	55.4	65

It is evident that 200 characters in size for the URL adequately covers the majority of the dataset, including more than 95% of the samples. Therefore, the vectorization procedure adopts this particular value. URLs longer than 200 characters are trimmed to this length, while URLs that are less than 200 characters are extended to 200 characters by padding. A phishing attack's use of a fake or misleading website address is illustrated visually in Figure 4.



Sign in

Email (phone for mobile accounts)

customer@amazon.com

Password Forgot your password?

············

Figure 4. Internet spoofing

In our dataset, we ensure uniform dimensionality in the training stage, regardless of their initial URL lengths. For URLs that are not longer than the specified dimension, the length is padded, while those longer than the prescribed length are shortened to conform. Every dataset entry is subject to the same uniform use of this URL length parameter. Choosing the optimal URL length involves finding a harmony between the architectural intricacy and the covering of the model. The length of the URL might greatly increase the amount of time needed for testing and training, thereby decreasing the effectiveness and usefulness of the model. Conversely, an overly short URL length compromises the model's training effectiveness. Therefore, for this study, a URL length of 200 was selected as it effectively encompasses the majority of the dataset. The histogram figure in Figure 7 illustrates the distribution of URL lengths in the 'big_dataset,' categorized into buckets of 20 characters each.

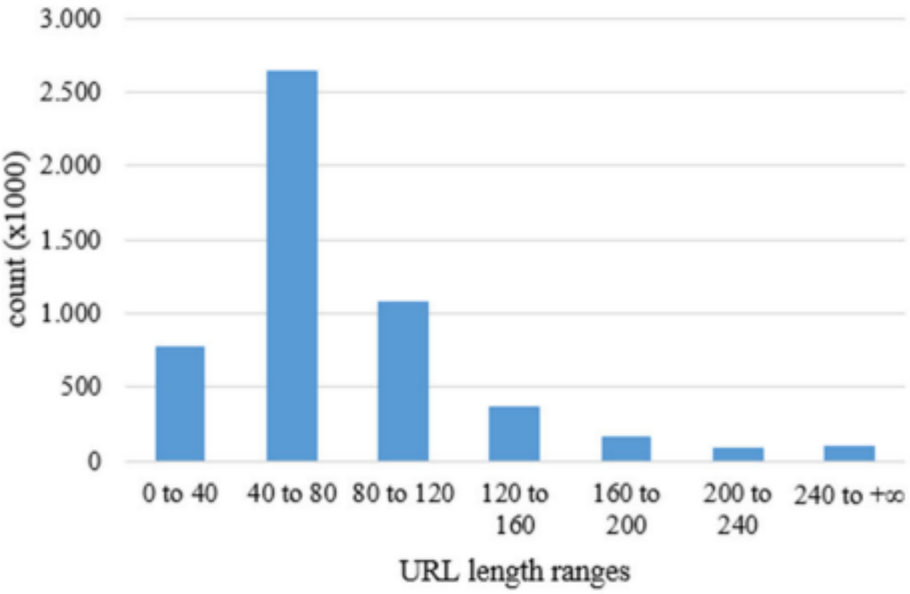


Figure 5. URL length range Histogram

Adversaries may seek to overcome security measures by using URL shorteners. Nonetheless, even if the service for verifying URLs installed in my research model fails to effectively classify the shortened URLs, the attackers' goals remain unattainable. This is due to the abbreviated URLs redirecting to the original web pages during the attack, ultimately being correctly categorized as phishing by the system.

This paper introduces a deep learning system for phishing URL detection, tailored for sequential data. Recognizing the importance of processing such sequences, the proposed model treats the URL as a cohesive entity, leveraging techniques commonly used in semantic analysis. Efforts were made to select algorithms favored in similar literature, balancing complexity with the need for swift response in cyber-attack detection. Five distinct deep learning architectures were evaluated, with configurations explored to enhance accuracy.

Artificial Neural Networks (ANN) employs acyclic connections between units, primarily dense layers. Convolutional neural networks (CNNs) require minimal preprocessing and excel at identifying patterns within sentences, making them suitable for word-based approaches. For character-based methods, CNNs effectively recognize malicious patterns in phishing URLs.

Recurrent Neural Networks (RNN) is adept at handling sequential data, processing each word with knowledge of preceding words. This architecture facilitates information flow through a chain-like structure, ideal for analyzing sequential data.

3. Results

This research project involved the development of a deep learning system aimed at identifying phishing URLs. The system's performance was evaluated by comparing it with conventional machine learning techniques. Deep learning models were constructed utilizing the Keras and TensorFlow frameworks, while traditional algorithms were assessed using Scikit-Learn.

Evaluation metrics such as accuracy, loss, precision, recall, and F-score were employed, alongside considerations of execution time on an Nvidia Tesla V100 GPU-equipped server. Hyperparameters for deep learning algorithms, including loss function, sequence size, and character embedding dimension, were determined, with default values used for others.

Figure 6 illustrates, five architectures for deep learning: ANNs, CNNs, RNNs, BRNNs, and ATTs, with batch size fine-tuned to optimize GPU utilization. Tests were conducted using a standard pattern of a single layer of 128 neurons, followed by performance enhancement techniques. Results were analyzed over 20 epochs.

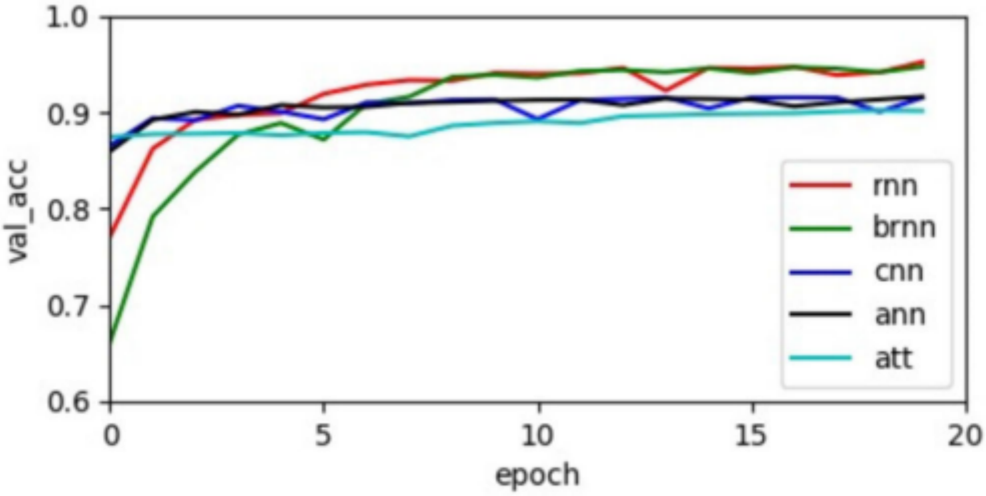


Figure. 6 Accurate base scenario validation

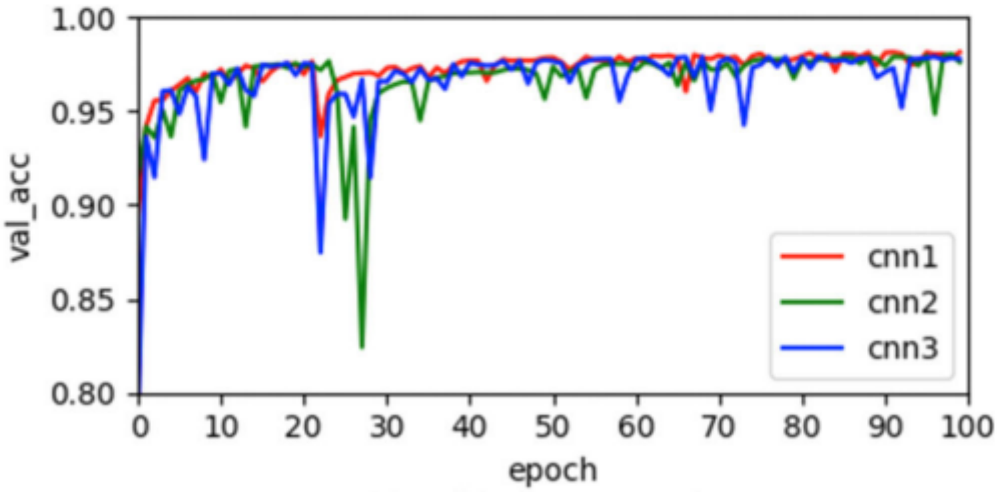


Figure 7. CNN validation precision

Figure 7 displays the performance metrics for the testing carried out on three intricate CNN architectures using smaller datasets. The tests carried out in the subsequent phase spanned 100 epochs.Two other CNN architectures were also looked into.

4. Discussion

This paper aims to develop a useful phishing detection solution that makes use of deep learning methods, leveraging a comprehensive dataset comprising over 2.3 million phishing URLs and nearly 2.9 million legitimate website URLs. Deep learning algorithms, chosen for their ability to handle large datasets, were evaluated using an Nvidia Tesla V100 GPU. Two dataset versions,'small_dataset' and 'big_dataset,' were provided to cater to diverse research needs and computational resources.

Various hyperparameters were fine-tuned to optimize deep learning model performance, initially tested on the 'small_dataset' and later validated on the larger 'big_dataset.' Among the architectures tested, convolutional neural network (CNN) showed the highest accuracy, achieving 98.74%.

Additionally, the system demonstrated impressive processing speed, classifying over 130,000 URLs per second, with the potential for further enhancement using more powerful processing units. Conventional methods for machine learning like Naïve Bayes, Random Forest, and Logistic Regression were also assessed alongside deep learning models.

The proposed CNN-based deep learning model offers notable advantages:

1. Language independence: It detects phishing attacks based solely on URL characters, regardless of language.
2. Extensive Dataset: Utilizes a vast training dataset for high accuracy.
3. Real-Time Execution: Processes URLs swiftly, suitable for deployment in high-traffic environments, with room for improvement using powerful GPUs.
4. Detection of New Websites: identifies new phishing sites, bolstering resilience against zero-day attacks.
5. Independence from Third-Party Services: Functions autonomously, ensuring efficient phishing detection without relying on external services, which is crucial for real-time operation, especially in high-traffic scenarios.

5. Conclusions

With the Internet's increasing importance, securing online assets becomes crucial. Phishing attacks exploit human vulnerabilities, posing a significant cyber threat. This paper presents a deep learning-based system to detect phishing attacks, evaluating five architectures: recurrent, bi-directional, convolutional, artificial, and attention-based neural networks.

The study shows deep learning outperforms traditional machine learning methods. It focuses on URL-based phishing detection, aiming for transparency and accessibility by sharing all codes and datasets accessible via IEEE DataPort, Code Ocean, and Github. Detailed test data is one of these resources, guaranteeing repeatability and an in-depth examination of the results.

References

1. A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab, “A comprehensive survey for intelligent spam email detection,” IEEE Access, vol. 7, pp. 168261–168295, 2019, doi: 10.1109/ACCESS.2019.2954791.
2. L. Tang and Q. H. Mahmoud, “A Deep Learning-Based Framework for Phishing Website Detection,” IEEE Access, vol. 10, pp. 1509–1521, 2022, doi: 10.1109/ACCESS.2021.3137636.
3. F. Sattari, A. H. Farooqi, Z. Qadir, B. Raza, H. Nazari, and M. Almutiry, “A Hybrid Deep Learning Approach for Bottleneck Detection in IoT,” IEEE Access, vol. 10, no. July, pp. 77039–77053, 2022, doi: 10.1109/ACCESS.2022.3188635.
4. Y. A. Alsariera, V. E. Adeyemo, A. O. Balogun, and A. K. Alazzawi, “AI Meta-Learners and Extra-Trees Algorithm for the Detection of Phishing Websites,” IEEE Access, vol. 8, pp. 142532–142542, 2020, doi: 10.1109/ACCESS.2020.3013699.
5. A. El Aassal, S. Baki, A. Das, and R. M. Verma, “An In-Depth Benchmarking and Evaluation of Phishing Detection Research for Security Needs,” IEEE Access, vol. 8, pp. 22170–22192, 2020, doi: 10.1109/ACCESS.2020.2969780.
6. S. Poudyal and Di. Dasgupta, “Analysis of Crypto-Ransomware Using ML-Based Multi-Level Profiling,” IEEE Access, vol. 9, pp. 122532–122547, 2021, doi: 10.1109/ACCESS.2021.3109260.
7. S. Anwar et al., “Countering Malicious URLs in Internet of Things Using a Knowledge-Based Approach and a Simulated Expert,” IEEE Internet Things J., vol. 7, no. 5, pp. 4497–4504, 2020, doi: 10.1109/JIOT.2019.2954919.
8. Z. Azam, M. M. Islam, and M. N. Huda, “Comparative Analysis of Intrusion Detection Systems and Machine Learning-Based Model Analysis Through Decision Tree,” IEEE Access, vol. 11, no. August, pp. 80348–80391, 2023, doi: 10.1109/ACCESS.2023.3296444.
9. M. Aljabri et al., “Detecting Malicious URLs Using Machine Learning Techniques: Review and Research Directions,” IEEE Access, vol. 10, no. October, pp. 121395–121417, 2022, doi: 10.1109/ACCESS.2022.3222307.
10. J. Nicholls, A. Kuppa, and N. A. Le-Khac, “Financial cybercrime: A comprehensive survey of deep learning approaches to tackle the evolving financial crime landscape,” IEEE Access, vol. 9, pp. 163965–163986, 2021, doi: 10.1109/ACCESS.2021.3134076.
11. C. Zonyfar, J.-B. Lee, and J.-D. Kim, “HCNN-LSTM: Hybrid Convolutional Neural Network with Long Short-Term Memory Integrated for Legitimate Web Prediction,” J. Web Eng., vol. 22, pp. 757–782, 2023, doi: 10.13052/jwe1540-9589.2251.
12. A. Raza, K. Munir, M. S. Almutairi, and R. Sehar, “Novel Class Probability Features for Optimizing Network Attack Detection With Machine Learning,” IEEE Access, vol. 11, no. September, pp. 98685–98694, 2023, doi: 10.1109/ACCESS.2023.3313596.
13. Y. Fang, C. Zhang, C. Huang, L. Liu, and Y. Yang, “Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism,” IEEE Access, vol. 7, pp. 56329–56340, 2019, doi: 10.1109/ACCESS.2019.2913705.
14. M. Sanchez-Paniagua, E. F. Fernandez, E. Alegre, W. Al-Nabki, and V. Gonzalez-Castro, “Phishing URL Detection: A Real-Case Scenario Through Login URLs,” IEEE Access, vol. 10, pp. 42949–42960, 2022, doi: 10.1109/ACCESS.2022.3168681.
15. T. Chin, K. Xiong, and C. Hu, “Phishlimiter: A Phishing Detection and Mitigation Approach Using Software-Defined Networking,” IEEE Access, vol. 6, pp. 42513–42531, 2018, doi: 10.1109/ACCESS.2018.2837889.
16. G. Lin, N. Sun, S. Nepal, J. Zhang, Y. Xiang, and H. Hassan, “Statistical Twitter Spam Detection Demystified: Performance, Stability and Scalability,” IEEE Access, vol. 5, pp. 11142–11154, 2017, doi: 10.1109/ACCESS.2017.2710540.
17. S. Razaulla et al., “The Age of Ransomware: A Survey on the Evolution, Taxonomy, and Research Directions,” IEEE Access, vol. 11, no. April, pp. 40698–40723, 2023, doi: 10.1109/ACCESS.2023.3268535.
18. J. Feng, L. yang Zou, O. Ye, and J. zhou Han, “Web2Vec: Phishing Webpage Detection Method Based on Multidimensional Features Driven by Deep Learning,” IEEE Access, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3043188.
19. K. M. Karthick Raghunath, V. Vinoth Kumar, M. Venkatesan, K. K. Singh, T. R. Mahesh, and A. Singh, “XGBoost Regression Classifier (XRC) Model for Cyber Attack Detection and Classification Using Inception V4,” J. Web Eng., vol. 21, no. 4, pp. 1295–1322, 2022, doi: 10.13052/jwe1540-9589.21413.
20. A. D. C. Manoj Kumar Prabakaran, Parvathy Meenakshi Sundaram, “An enhanced deep learning-based phishing detection mechanism to effectively identify malicious URLs using variational autoencoders”, doi: https://doi.org/10.1049/ise2.12106.

21. M. S.-P. E. F. F. E. A. W. Al-Nabk, “Phishing URL Detection: A Real-Case Scenario Through Login URLs”, doi: 10.1109/ACCESS.2022.3168681.

22. A. S. Das Guptta, K. T. Shahriar, H. Alqahtani, D. Alsalman and ‘ I. H. Sarker, “Modeling hybrid feature-based phishing websites detection using machine learning techniques,’’,” 2022, doi: 10.1007/s40745-022-00379-8.

23. J. Anitha and M. Kalaiarasu, “A new hybrid deep learning-based phishing detection system using MCS-DNN classifier,” 2022, doi: 10.1007/s00521- 021-06717-w.

24. and W. W. Y. Huang, Q. Yang, J. Qin, “Phishing URL detection via CNN and attention-based hierarchical RNN,” 2019, doi: 10.1109/Trust- Com/BIGDATASE.2019.00024.