STM JOURNALS

Journal of
Computer Technology & Applications

ISSN: 2229-6964 (Online)
ISSN: 2347-7229 (Print)
Volume 15, Issue 3, 2024
September–December
DOI (Journal): 10.37591/JoCTA

https://journals.stmjournals.com/jocta

Research JoCTA

Detecting Phishing Websites Using Hybrid Methodologies

Gargi Deshpande[1,*], Shreyas Katkar[2], Tanvi Kangane[3], Ayush Kumar Giri[4], Diksha Kale[5]

Abstract

In the digital era, personal information theft has become a widespread and increasingly severe crime. Cybercriminals, often known as hackers, use deceptive strategies, with phishing websites being a major method for stealing confidential data. These fake websites imitate legitimate ones, tricking users into revealing sensitive personal and financial information, which has led to a rise in fraud cases. To address this escalating threat, a comprehensive research paper is proposed. This approach involves preprocessing datasets and analyzing attributes such as Internet Protocol (IP) addresses, universal resource location (URL) length, and web traffic statistics to differentiate phishing websites from genuine ones. Feature extraction is performed using deep learning methods, such as convolutional neural networks (CNN) and gated recurrent units (GRU). A hybrid model that integrates machine learning with transformer and GRU components is used, demonstrating better performance than traditional methods like long short-term memory (LSTM), Naïve Bayes, and support vector machines (SVM). The study aims to identify phishing URLs and determine the most effective machine learning method by comparing each algorithm's accuracy, false positive rate, and false negative rate. The proposed method was evaluated using 156,422 malicious sites and 392,924 legitimate sites, and it was found to be more effective in detecting malicious URLs compared to more recent methods.

Keywords: Phishing, legitimate, CNN-GRU, LightGBM, length normalization, uniform encoding, embedding layer, SoftMax

INTRODUCTION

As our lives have become increasingly linked to the Internet, the risk of cyber-attacks, including phishing schemes, has surged significantly. The goal of phishers is to obtain vital data such as bank account details, usernames, and passwords. Phishing attacks entail the creation of phony websites or emails that imitate authentic websites to deceive users into disclosing sensitive data, including credit card details, passwords, and personal identification numbers. To fool users, phishers send as many "spoof" emails as they can to as many recipients. Customers who open these emails usually end up on a phone website instead of a real company. The impact of a successful phishing attack can be serious, leading to financial loss, identity theft, and even damage to one's reputation. Most phishing webpages have the same appearance and feel as legitimate webpages, down to the universal resource location (URL) and website interface. Heuristics and blacklisting are two techniques suggested for detecting phishing websites. However, owing to inadequate security measures, the number of victims has increased exponentially. The uncontrolled and anonymous nature of the Internet increases the

***Author for Correspondence**
Gargi Deshpande
E-mail: gargideshpande05@gmail.com

[1-4]Student, Department of Computer Engineering, Pillai College of Engineering, Dr. K.M. Vasudevan Pillai Campus, New Panvel East, Panvel, Navi Mumbai, Maharashtra, India
[5]Assistant Professor, Department of Computer Engineering, Pillai College of Engineering, Dr. K.M. Vasudevan Pillai Campus, New Panvel East, Panvel, Navi Mumbai, Maharashtra, India

Received Date: June 26, 2024
Accepted Date: August 12, 2024
Published Date: September 12, 2024

Citation: Gargi Deshpande, Shreyas Katkar, Tanvi Kangane, Ayush Kumar Giri, Diksha Kale. Detecting Phishing Websites Using Hybrid Methodologies. Journal of Computer Technology & Applications. 2024; 15(3): 59–65p.

frequency of phishing attacks. Previous studies have indicated that the performance of phishing detection systems is constrained. A clever strategy is required to safeguard users against cyber-attack.

Cybersecurity professionals are currently searching for reliable and consistent methods to identify phishing websites. To guarantee a reliable and thorough approach to the detection of phishing URLs, this study combined many cutting-edge approaches by extracting and evaluating different aspects of legal and phishing URLs. This multifaceted approach encompasses feature-based analysis, machine learning (ML), and visual similarity analysis, all of which are supported by cutting-edge technologies.

Convolutional neural networks (CNNs) have been employed for feature-based analyses. They are highly proficient at detecting complex patterns and features in data, making them particularly well-suited for identifying visual cues and anomalies in web pages. Gated recurrent units (GRU) play a crucial role in the machine learning aspect of the approach. GRUs are adept at modeling sequential data, making them invaluable for analyzing the dynamic behavior of websites over time. Additionally, LightGBM, a gradient-boosting framework, was used to improve the model's performance and accuracy. LightGBM excels in handling large datasets and can efficiently handle a wide range of features, making it well-suited for the complex task of differentiating between legitimate and phishing websites.

Objectives

The primary objective of this project was to develop a robust system for detecting phishing websites in real-time using a combination of standardized datasets from Kaggle and Phish Tank.

To accomplish this objective, new deep learning algorithms will be implemented to improve the accuracy and efficiency of the system.

Once deep learning algorithms have been developed, they will be integrated into the system, allowing the analysis of websites in real-time. This real-time analysis will give users instant feedback on the websites they are visiting, helping them make informed decisions about whether to continue.

Scope

The scope of the "Detecting Phishing Websites using Hybrid Methodologies" project is to develop a system that can accurately detect phishing websites. The project will use standardized datasets and implement new deep learning algorithms to analyze various aspects of a website and determine whether it is legitimate or a phishing website. The project also involves the development of a user-friendly interface to provide clear feedback to users. The main goal of the project is to develop an effective tool for preventing phishing attacks and safeguarding users from financial losses, identity theft, and damage to their reputation.

RELATED WORK

This section elaborates on the research carried out on some of the existing systems.

Existing Research

Yang *et al.* [1] proposed a system in which the model was divided into three major algorithms: CNN-long short-term memory (LSTM), DCDA, and multidimensional features. The CNN-LSTM algorithm includes URL-embedded representation, feature extraction, and classification. Normalization of the URL character sequence is performed, and it is converted to a one-hot code sequence, which is then converted to an embedding matrix. From this embedding matrix, features were extracted by the CNN layers. This result is provided as an input to the LSTM. The LSTM output is the input for the SoftMax unit, which uses the dropout method to prevent overfitting. The Dynamic Category Decision algorithm is a probabilistic method that improves the SoftMax layer output. Based on the results, it is decided whether to pass the URL to the next step. The multidimensional feature algorithm extracts static URL features, such as webpages and text features, which are assigned to the XGBoost classifier, which classifies the URL as phishing or legitimate.

Odeh *et al.* [2] researched different challenges and proposed techniques. Here are the ideas reviewed in the paper:

1. Deep learning approach to overcome the limitations of heuristic search and Random Forest (RF).
2. We used a real Internet Protocol (IP) from an Internet Service Provider (ISP) and Deep Belief Networks (SBN) to obtain the results.
3. Data-mining techniques can be used to obtain a large amount of data. Machine learning and feature extraction can be applied to classify the data as phishing, suspicious, or legitimate.
4. Machine learning anomalies and attacks are used to capture how sensitive information is obtained by hackers.
5. The stacking model method was used to detect legitimate websites, and it provided higher accuracy than many ML models.
6. Machine learning with wrapper-based features yielded more promising results than some benchmark models.
7. The combined method of using a machine learning classifier and wrapper-based feature selection worked better in detecting phishing websites.
8. A deep learning approach that does not require any prior knowledge and uses URL character sequencing for quick classification.
9. The most suitable ML algorithms were identified: K-nearest neighbor (KNN), SVM classifier, Random Forest, D-tree, linear-SVC classifier, and wrapper-based (W-B) feature selection. The SVM and Random Forest models performed more efficiently than the other models.
10. SVM-KNN hybrid approach using specific features.

Yi Wei and Yuji Sekiya [3] compared a list of machine learning models for various parameters, such as accuracy, precision, recall, and F1 score. The results were obtained by applying classical machine learning techniques such as Random Forest, logistic regression (LR), linear discriminant analysis (LDA), K-means clustering (k-means), support vector machine (SVM), Naïve Bayes classifier (NB), KNN, and LR. Among these algorithms, RF gave the highest accuracy and k-means the least. In addition, ensemble ML methods include gradient-boosted decision trees (GBDT), AdaBoost, XGBoost, LightGBM, and RF. RF yielded the highest accuracy, and AdaBoost the least. The study also includes deep learning-based methods such as fully connected neural networks (FCNN), LSTM, and CNN. These models were compared with RF where RF achieved the highest accuracy.

Bhagwat *et al.* [4] divided their work into two parts: detection and prevention. To detect whether the website is phishing or genuine, the URL is pre-processed based on various website features, and the ML algorithm is applied. The output is passed to the Fuzzy Inference System, which uses a rule-based method to obtain the phishing rate based on which the URL is classified. For the prevention part, after the website is detected as phishing based on the phishing rate, the website's IP address, host location, and host server information are fetched so that the administrator can remove that page.

Zhou *et al.* [5] approached this problem by extracting domain-based features from a labeled dataset. This preliminary extraction of domain information was provided as an input for feature selection. Here, one-hot encoding, normalization variance selection, is performed on the data. The features are selected such that they improve the accuracy of the model. This is further provided as an input to the LightGBM classification model, which categorizes the URL as a normal or phishing website. Based on the experimental results, it was concluded that the proposed model, which uses two types of features for training, yields a better outcome than the model that uses only a single feature.

Alshingiti *et al.* [6] used a deep learning approach. The URLs are pre-processed, including data cleaning, data transformation using the MinMax scaler, and feature engineering using SelectKBest. This result is given to the three deep learning models (keeping the dataset the same): CNN, LSTM, and LSTM-CNN, separately. The final step includes the classification of URLs as phishing or legitimate. The models were evaluated based on their precision, recall, accuracy, and F1 scores. The results indicated that the CNN model performed best across all evaluation metrics.

Research Gap

After conducting a literature survey on the existing systems, it was noticed that the latest machine learning and deep learning models have not been implemented, which will succeed in providing more accurate results and consuming less time. In addition, as new malicious websites are created, it is important to train the model using the latest datasets [7].

METHODOLOGY

This section explains the dataset used, the proposed system, and various algorithms used.

Dataset

The dataset investigated in this study was obtained from Kaggle and served as the foundation for our research. It comprises a total of 5,49,346 records and has two columns: URL and Label. The URL column contains the URLs of the sites, whereas the label column has two categories.

- *Good:* Websites without harmful content (not phishing sites).
- *Bad:* Websites with malicious content (phishing sites).

There were no missing values in the dataset. The dataset was meticulously explored to extract meaningful patterns, train the predictive models, and draw insightful conclusions.

Proposed System

The goal of this study was to develop a hybrid phishing detection approach. Instead of relying on a single technique, we integrated multiple methods. This allows us to leverage their strengths and offset their weaknesses [8]. The result is a more confident way to identify phishing websites [9].

In the first step, before applying the CNN-GRU model, the URL character sequences were pre-processed. Length normalization ensures that all the URL sequences have a fixed size [10]. If a URL is too long (exceeding a certain length, denoted as "L"), the extra characters are trimmed from the end. If a URL is too short (less than "L"), zeros are added until it reaches the desired length. Subsequently, we performed uniform encoding, that is, we converted the normalized URL string into a one-hot code sequence. This encoding ensures that the URL is presented in a uniform format for the subsequent analysis. Finally, we pass it through the embedding layer, and to enhance the efficiency, we transform the sparse one-hot matrix into a denser character embedding matrix. This step obtains the data ready for the machine learning algorithms.

Once the URL data are prepared, essential features are extracted from the character sequences and used for classification. A CNN captures local correlations within the URL, with the convolutional layer identifying patterns and the pooling layer reducing the dimensionality. Additionally, a GRU network was employed to understand the context dependencies in the URL. Finally, SoftMax is used for classification, assigning probabilities to different classes (e.g., "Good" or "Bad" URLs).

In the second step, after ensuring that the URL is accessible, we extracted a combination of features related to the address bar, domain, HTML, and JavaScript. Address bar based features include domain details, IP address in URL, "@" symbol presence, use of URL shortening services "Tiny URL," "http/https" in the domain name, URL length, redirection "//" in URL, and more. Domain-based features include DNS records, website traffic, domain age, and domain expiration date. Features, such as website forwarding and I-frame redirection, are based on HTML and JavaScript. In addition, we incorporated the classification results from our deep learning model (CNN-GRU). Finally, we utilized LightGBM for the ultimate classification. The deep features extracted from the CNN-GRU algorithm serve as crucial inputs for this process. If the URL cannot be accessed or is unreachable, we rely directly on the classification result obtained from the SoftMax output in the initial step. In other words, SoftMax prediction serves as our detection outcome when direct URL access is not possible.

Algorithm Used

CNN

A CNN is a form of deep learning architecture used to extract local features from URLs. The CNN framework includes three layers: a convolution layer, a pooling layer, and a fully connected layer. The convolution layer extracts various features from the input, whereas the pooling layer reduces the size (or spatial extent) of convolved features. The fully connected (FC) connects neurons between layers and processes flattened features (Figure 1).

GRU

The GRU is a form of recurrent neural network (RNN) used in deep learning and was introduced in 2014. Although it shares characteristics with LSTM networks, it functions as a more straightforward substitute for LSTM networks. The GRU is capable of handling sequential data, including speech, text, and time-series data, similar to LSTM. The advantage of GRU lies in its reduced number of training parameters, resulting in lower memory usage and faster execution compared to LSTM.

LightGBM

LightGBM is a gradient-boosting framework designed to reduce memory usage and enhance model efficiency. It is commonly used for ranking, classification, and various other machine learning tasks. Unlike other boosting algorithms, such as XGBoost, LightGBM performs tree splits leafwise rather than levelwise. Its outstanding features include higher accuracy, faster training speed, low memory footprint, and compatibility with both small and large datasets.

RESULTS

This section discusses the various evaluation parameters used to analyze the results in Table 1.

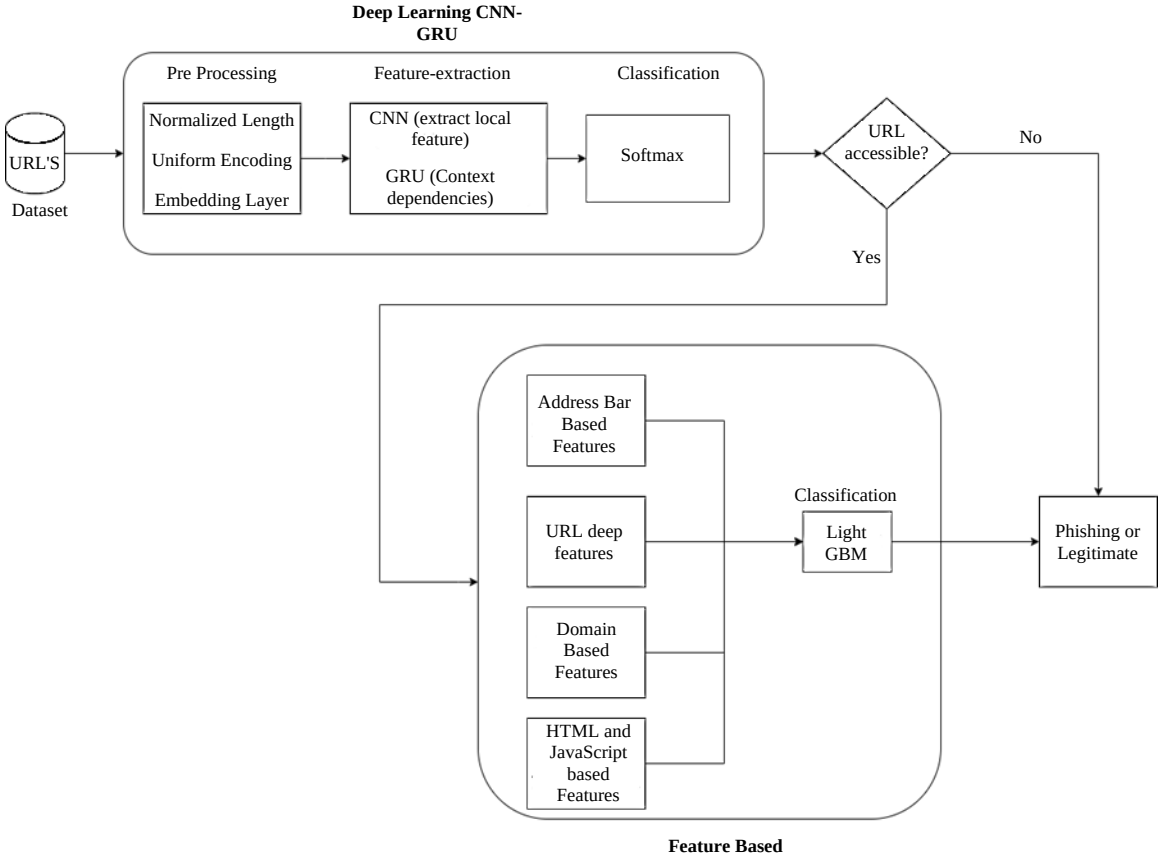

Figure 1. Proposed system architecture.

Table 1. Implementation results.

Evaluation parameter	GRU	LightGBM
Accuracy	97.43	96.23
Precision	97.31	93.42
Recall	98.71	99.13
F1 Score	98.01	96.05

Accuracy

These metrics measure the overall accuracy of a model's predictions. The GRU achieved an accuracy rate of 97.43%, indicating that 97.43% of the predictions made by the model were accurate. Conversely, LightGBM exhibited a slightly lower accuracy rate of 96.23%.

Precision

Precision is the ratio of the model's total number of successful predictions to the number of real positive forecasts. In this situation, GRU achieved a precision of 97.31%, suggesting that 97.31% of the websites classified as phishing were indeed phishing. Meanwhile, LightGBM demonstrated a precision of 93.42%.

Recall

This quantifies the proportion of actual positives in the dataset to the total number of true positive predictions. GRU recorded a recall rate of 98.71%, indicating that 98.71% of actual phishing websites were correctly identified by the model. On the other hand, LightGBM exhibited a slightly superior recall rate of 99.13%.

F1 Score

The F1 score provides an accurate assessment of a model's performance by being the harmonic mean of the precision and recall. GRU achieved an F1 score of 98.01%, whereas LightGBM achieved an F1 score of 96.05%.

CONCLUSION AND FUTURE WORK

In summary, our study offers a new strategy for identifying phishing websites by integrating a hybrid methodology that combines GRU and deep learning with LightGBM, a machine learning algorithm. By leveraging various attributes, such as IP addresses, URL length, and web traffic statistics, our system aims to empower users with the ability to make informed decisions while navigating the vast online landscape. Our analysis demonstrates that the hybrid model outperforms individual deep learning or machine learning algorithms, demonstrating its efficacy in accurately distinguishing between legitimate and phishing websites.

As online threats become increasingly sophisticated, it is crucial to enhance security measures to safeguard users from potential cybercrime. Our proposed system represents a significant contribution to this endeavor by providing a robust framework for website authentication to enhance the overall security posture of Internet users, ultimately fostering a safer online environment.

To improve the system, the following features can be considered:

- Explore more advanced deep learning architectures and feature engineering techniques.
- Incorporate larger and more diverse datasets for better model generalization.
- Integrate user behavior analysis for enhanced phishing recognition.
- Create systems that adapt to changing attack strategies.

REFERENCES

1. Yang P, Zhao G, Zeng P. Phishing website detection based on multidimensional features driven by deep learning. IEEE Access. 2019;7:15196–209. DOI: 10.1109/ACCESS.2019.2892066.

2. Odeh A, Keshta I, Abdelfattah E. Machine learning techniques for detection of website phishing: A review of promises and challenges. In: 11th Annual Computing and Communication Workshop and Conference (CCWC); 2021. Las Vegas, USA; 0813–818.
3. Wei Y, Sekiya Y. Sufficiency of ensemble machine learning methods for phishing websites detection. IEEE Access. 2022;10:124103–13. DOI: 10.1109/ACCESS.2022.3224781.
4. Bhagwat MD, Patil PH, Vishawanath TS. A Methodical Overview on Detection, Identification and Proactive Prevention of Phishing Websites. In: Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV); 2021. Coimbatore, India; 1505–1508. DOI: 10.1109/ICICV50876.2021.9388441.
5. Zhou J, Cui H, Li X, Yang W, Wu X. A novel phishing website detection model based on LightGBM and domain name features. Symmetry. 2023;15:180. DOI: 10.3390/sym15010180.
6. Alshingiti Z, Alaqel R, Al-Muhtadi J, Haq QEU, Saleem K, Faheem MH. A deep learning-based phishing detection system using CNN, LSTM, and LSTM-CNN. Electronics. 2023;12:232. DOI: 10.3390/electronics12010232.
7. Alswailem A, Alabdullah B, Alrumayh N, Alsedrani A. Detecting phishing websites using machine learning. In: 2nd International Conference on Computer Applications & Information Security (ICCAIS); 2019. 1–6. DOI: 10.1109/CAIS.2019.8769571.
8. Xu Y, Chen G, Liu Q, Xu W, Zhang L, Wu J, Fan X. A phishing website detection and recognition method based on naive Bayes. In: 6th Information Technology and Mechatronics Engineering Conference (ITOEC); 2022. 1557–1562. DOI: 10.1109/ITOEC53115.2022.9734474.
9. Yuan H, Chen X, Li Y, Yang Z, Liu W. Detecting phishing websites and targets based on URLs and webpage links. In: 24th International Conference on Pattern Recognition (ICPR); 2018. 3669–3674. DOI: 10.1109/ICPR.2018.8546262.
10. Korkmaz M, Sahingoz OK, Diri B. Detection of phishing websites by using machine learning-based URL analysis. In: 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT); 2020. 1–7. DOI: 10.1109/ICCCNT49239.2020.9225561.

