Detecting Phishing Websites Using Hybrid Feature Extraction and Classification

D.Shunmuga Kumari,
Department of Computer Science and Information Technology, Vels Institute of science technology and advanced studies
Chennai, Tamil Nadu , India
kumari.vnr@gmail.com

S.Sathya
Department of Computer Science and Information Technology, Vels Institute of science technology and advanced studies,
Chennai, Tamil Nadu , India
ssathya.scs@vistas.ac.in

K.Prakash
Department of Computer Science,Sri Sarada Mahavidyalayam Arts and Science College for Women ,
Ulundurpet Tamil Nadu, India
prakash.staff@gmail.com

Sheela K
Department of Computer Science and Information Technology, Vels Institute of science technology and advanced studies
Chennai, Tamil Nadu , India
drksheela.research@gmail.com

M.Sakthivanitha
Department of Computer Applications,Vels Institute of science technology and advanced studies Chennai, Tamil Nadu , Chennai
sakthivanithamsc@gmail.com

V.Poornima
Department of Computer Science and Information Technology, Vels Institute of science technology and advanced studies
Chennai, Tamil Nadu , India
poornimasudhaagar@gmail.com

Abstract: Phishing websites are a huge cyber security threat because they deceive people into divulging sensitive information, most of the time without being detected by traditional cyber security measures. The basic problem with phishing is that it can get very dynamic and change sometimes, with lexical tricks, bad content, or the deceiving characteristics of hosting, making it rather difficult to determine correctly and fairly if it's phishing or not. This research aims to present a hybrid extraction and classification framework to detect the occurrences of a phishing containing lexical, content-based, domain level (hybrid pattern) and behavioral features. Specifically, this research entails data pre-processing from the UCI Phishing Websites Dataset, hybrid feature engineering with dimensionality reduction, as well as ensemble methods, which include Random Forest, Gradient Boosting, Support Vector Machines, and a stacking meta-classifier. The quantitative results that are presented show that the hybrid model is 98.2% accurate compared to the current state-of-the-art methods such as CNN-LSTM and deep neural networks. Hybrid methods show a great impact in the accuracy in detection and still effective on emergent threats. In summary, the presented research identifies scalable and practical framework for real-time detection in regards to phishing and future work will be focused on adaptive deep learning models and explainable AI.

Keywords: Phishing Detection, Hybrid Feature Extraction, Ensemble Classification, Cybersecurity, Machine Learning, Website Security, Fraud Detection

I. INTRODUCTION

In the world of digital markets, the Internet is the basis of communication, commerce, and knowledge transfer. However, along with being useful, the Internet has also posed a situation where unethical individuals work to harm organisms and infect systems by exploiting vulnerabilities for fraud, identity theft and crimes relating to finance[1]. One of the most widespread and destructive types of cybercrime is phishing. Phishing is typically a trick on users by posing as an individual or business they trust to acquire sensitive credentials. Over the past several years, security tools and other practices have advanced, but phishing attacks have also evolved and use various tactics to attack users; these other tactics include domain spoofing, malicious scripting and AI-driven deception mechanisms. Phishing is said to be the cause for the majority of data breaches and billions of dollars lost. The consummate sophistication of phishing attacks requires intelligent detection systems that are capable of evolving with the phishing tactic[2].

Conventional phishing detection techniques mostly rely on the blacklist filtering method, which is based on identifying the malicious domains in advance. While techniques using the blacklist work well in limited situations, they fail to detect new or unknown phishing domains otherwise known as zero-day attacks. In order to overcome this problem, machine learning based detection techniques have been created using handcrafted features such as length of a URL, number of sub-domains and SSL certificate validity. The next generation of techniques incorporates deep learning, models based on Convolutional Neural Networks (CNN) and Long Short Term Memory (LSTM), which extract complex patterns from raw sources of information including URL strings or screenshots of a webpage[4]. Despite these advancements there are still limitations on techniques including computational cost, overfitting to a given dataset and lack of interpretability. Furthermore, majority of research, till now only focused on one category of phishing features, i.e., lexical, content or domain type features causing failure to capture many holistic characteristics of phishing attacks[5].

Research Gap and Problem Statement: while phishing detection research has made substantially moves, there continue to be limitations with respect to detection capabilities. Detection methods work with high accuracy on labelled datasets, but after that fail to generalize on real world settings, where the tactics of phishing attacks are constantly changing. Moreover, much of the foundation of phishing websites involve ignoring dynamic behavioral features that attackers are interested in using and phishing websites are using, such as redirecting users, anomalies in SSL handshakes, and abnormal or time delayed request-response times. THe gap is identified in the literature and shown in lack of converged frameworks to use different categories of features or in studies that use unadaptable classifiers. Therefore, the problem of research explores is as follows: How to use hybrid feature extraction and ensemble classification to improve the accuracy and adaptability of phishing website detection in real cases?

- To come up with a hybrid feature extraction framework comprising of lexical and content-based, domain-level and behavioral features of the web sites.
- To implement ensemble classification strategy along with the integration of multiple machine learning models using a stacking meta-classifier for more robustness.
- To set forth to compare the performance of the proposed model with state-of-the-art methods for detecting phishing attacks (using public benchmarks).

The rest of this paper is organized as follows: Related work in the area of phishing detection is reviewed in Section 2. Section 3 describes the proposed methodology based on both feature extraction and classification. Section 4 is the description of the dataset and the experimental results. The discussion of the findings, the limitations and the implications is in Section 5. The study, and possible future areas of research, will be concluded on Section 6.

II. RELATED WORKS

Recent studies that have developed phishing detection have made use of deep learning methods, ensemble techniques, and feature selection efforts for effectiveness and improvement in accuracy. For instance, some studies show a transition from the traditional methods of detecting phishing attacks, while other important studies show progressions in the use of GANs-generated URLs, single or hybrid networks of neural networks, stacked ensembles, feature selection, semantic-based analysis, and multimodal embeddings, which were shown to be superior to traditional and also to provide more than 95% detection accuracy across the several datasets and even in disparate scenarios.

Said et al. (2024) have developed a phishing detection application based on CNN + self attention mechanism treating URLs as normalized strings. The URLs GAN produced led to a balanced model throughout the training process. The model performed with precision of 99.7% surpass standard CNN's performed on previously unseen URLs, and ultimately, improved accuracy receipts for phishing detection improvements[6].

Kalabarige et al. (2022) proposed a multi-layered stacked ensemble learning approach to phishing website detection, where the predictions of one layer are used to predict the next layer to improve the performance. Testing on the UCI and Mendeley data sets resulted in accuracy measures between 96.79 to 98.90% significantly outperforming baseline models in both tournaments in terms of accuracy and the F-score[7].

Moedjahedy et al. (2022) have proposed the CCrFS (correlation based recommender for feature selection), a feature selection technique that applied correlation measures combined with recursive feature elimination for detecting phishing sites. The model tested in data sets with 48, and 87 features were able to create impressive testing performance of between 97.06% to 95.88% for site detection while only using 10 features when selected[8].

Almomani et al. (2022) carried out a comparison study for the detection of phishing by using semantic features (URL, domain identity, abnormal, HTML/JavaScript, and domain attributes). They tested 16 classifiers, used by two different datasets, and found out that Gradient Boosting and Random Forest provided the highest accuracy (~97%) and that compared to other classifiers, such as GaussianNB (84%) and SGD (81%)[9].

Rao et al. (2022) designed a phishing detection technique which focused on word embeddings from both plain and domain-specific source code text. They tested the technique using ensemble and multimodal approaches and reported the achieved accuracy of 99.34% using multimodal approach with 99.59% TPR, 0.93% FPR, and 98.68% Matthews Correlation Coefficient (MCC) which meant extraordinary phishing detection[10] capability.

Das Gupta et al. (2024) proposed a hybrid feature-based phishing detection model based on both URL features and hyperlink features that do not require third-party (e.g., Google) for the web crawling, and allow real-time applicability. They came up with their own dataset for the phishing detection and have decided to test its objectives on the basis of accuracy. They found that XGBoost achieved an accuracy of 99.17%, which is better than those traditional methods which use blacklists, heuristic technique or visual similarity methods[11].

Kasim (2021) has provided an event-based phishing detection method based on deep-hybrid feature extraction using SAE-PCA and classification using Light GBM. For this purpose, they showed an example application of the method by using the ISCX-URL dataset and achieved an accuracy rating of 99.6% while presenting a reduction of false positives by detecting malware if it is first entering the web service request but before loading the entire source webpage[12].

Wen et al. (2023) proposed LBPS, which is a hybrid deep neural network that incorporates BP neural network and LSTM-FCN for Phishing scams identification of Ethereum accounts. The authors used transaction records to reach some impressive performance, including an F1-score of 97.86%. In comparison, LBPS had a significant improvement over the baseline approaches[13].

Even though the previous studies demonstrated an impressive performance, they still have many limitations or drawbacks. For example, several of these projects are reliant on certain data sets (e.g., UCI, Mendeley, ISCX-URL, Ethereum) that fail to capture the dynamic nature of phishing, which restricts the transferability of the research to the world. While the use of GANs to generate synthetic URLs helps to address the balance of data, the synthetic URLs created might not be able to accurately represent adversarial URLs. Selecting features to model may limit one's potential to study codependent or subtle features that may be important. Although ensemble and hybrid models increase the performance of the models, these models have increased their complexity and computing cost, which hinders their adoption and utility in systems that are designed to combat phishing in real-time. There were few if any, studies which studied zero-day phishing, whether a model approaches transferred across contexts, or research on phishing in languages other than English. Finally, most of the approaches studied here did not have a particular focus on explainability, and therefore it was difficult for decision-makers, such as security analysts, to understand and appropriately interpret the decisions taken by the model. Moving forward, studies should focus on how to improve scalability, adaptability, explainability and/or robustness against adversarial phishing attacks.

III. METHODOLOGY

The proposed method uses the data acquisition, hybrid feature extraction and ensemble classification for the identification of phishing websites as illustrated in figure 1. The collected data is then preprocessed in a uniform manner and different lexical, content, domain and behavioral features are extracted after that. Dimensionality reduction helps with the efficiency and the stacking of ensemble is useful toward the robustness, accuracy and adaptability for real-time application of phishing detection.

A. Data Acquisition & the Preprocessing

The proposed method entails systematic collection of data of web sites from publicly available phish databases, as the starting point for the process. To ensure a good quality, the raw data collected will be processed before being collected in order to remove duplicates, inactive links, as well as the standardization of the URL formats. Next, I will normalize the text in the web pages that have been collected and impute any missing values of the attributes using a nearest neighbor method. I will create variable encoding for features such that I will have a Homogenous encoding for features, for categorical variables, one-hot encoding approach will be used, and for continuous features, min max normalization will be performed. This will be helpful in providing a standard format of the lexical and behavioural features during the learning tasks in order to reduce noise and increase confidence in classification. The pre-processing pipeline is meant to be dynamical, allowing for updates to be made as new phishing methodologies are introduced, allowing the system to be in a state of flux. This consistent format for the dataset is the basis in the hybrid feature extraction which balances the robustness, generalizability and scalability for phishing websites detection in the real-world.

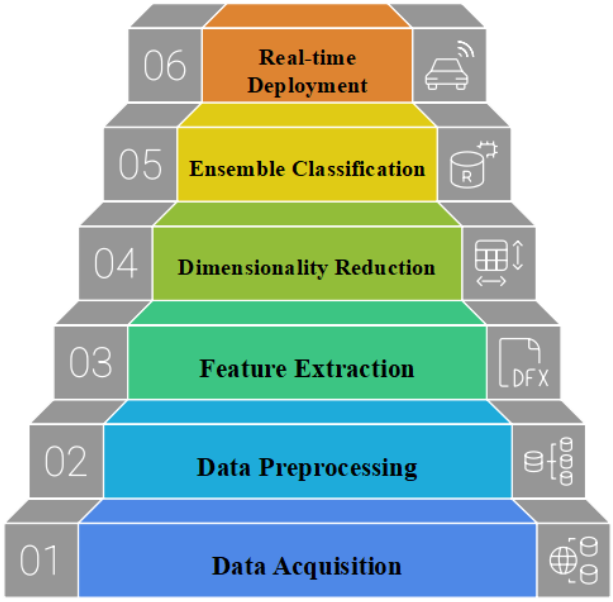

Figure 1 Steps in Phishing Detection

B. Hybrid Feature Extraction

To obtain multidimensional elements of a phishing website, a hybrid feature extraction framework is adopted. Lexical features are calculated based on URL pattern, length, number of special characters, entropy and presence of suspicious keywords. Concurrently, content-based features are extracted from the analysis of the source-code of the Web page (including the use of iframe, obfuscation patterns, and frequency of redirection) in the case of the Web page, and the JavaScript file texts. In addition to this, domain-based features such as WHOIS information, age of domain and DNS records are also included to detect abnormalities in the ownership and hosting behaviors. Moreover, behavioral features, such as validity of an SSL certificate, request-response time and behaviors of embedded links are included to add a further layer of detection. These different feature categories are joined up into a composite vector representation. To reduce the issues related to a redundancy of features, the work uses Principal Component Analysis (PCA) and Mutual Information ranking to reduce dimensionality, so as to remain efficient and preserving discriminative capacity. This hybrid extraction provides a combination of static and dynamic cues of phishing websites that develop a holistic detection mechanism.

C. Classification Framework

For classification, a hybrid ensemble learning strategy is applied in order to make use of complementary advantages of several algorithms. Figure 2 shows the framework of classification adopted in this study.

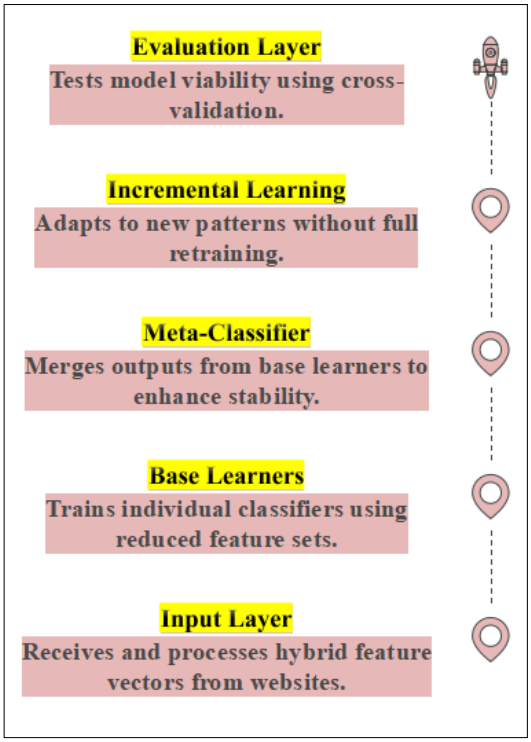

Figure 2 Robust Classification Frameworks

Input Layer: The input layer takes the processed hybrid feature vector which is the lexical-level, content-based level, domain-level, and behavior of websites. The features presented to the model include rich display of static and dynamic cues of phishing so that a holistic view of patterns (static and dynamic) used by an attacker can be obtained. Additionally, the reduced dimensionality (as a result of PCA and mutual information ranking) ensures the models are efficient and at the same time have the capacity of discrimination.

Base Learners : In the first part of the single phase two stage learning approach, three individual classifiers will be trained: Random Forest, Gradient Boosting and Support Vector Machine classifiers based on the reduced feature set. Each of the three algorithms have an even greater unique contribution factors relating to their optional approaches. Random Forest can be properly used for external - yet related - non-linear relationships. Gradient Boosting works to look into the error correction to increase the accuracy of the current and future external-yet-related subsets. Last, the SVM is maximizing class boundaries. Hyperparameters for the model will be optimized to be able to observe a complete picture of precision, recall, and generalization with hyperparameter optimization.

Meta-Classifer : The outputs from the base learners are combined using stacking framework which uses Logistic Regression as the meta-classifier. In this phase, the best way to combine the predictions is learned using the predictions of the different classifiers, while reducing the bias and variance of an individual classifier model. In addition, the meta-classifier is employed as a second level in decision fusion in order to increase the stability and robustness of the predictions.

Incremental Learning Module: To add to the adaptability of the application in the real world, we have built an online incremental learning module into the framework. If there are new phishing patterns or new or changing attack strategies discovered then there will be an incremental increase in the weights of the model without a full model retrains. One of the many advantages of the ongoing learning in our framework is that because it evolves, it makes our system more resilient to change and efficient for the long term.

Evaluation Layer: The evaluation layer is the layer used to test the viability of the model with stratified k-fold cross validation which gives the statistical reliability needed to evaluate the model at this stage. The performance of the hybrid classification systems is measured by different metrics, such as Accuracy, Precision, Recall, F1-Score, and AUC, to fairly reflect the classification performances of the classification system. Through this evaluation, the viability, responsiveness, and applicability for the real world of the hybrid classification framework will be evaluated.

IV. RESULTS AND FINDINGS

The results section gives the evaluation of the proposed hybrid phishing detection framework based on the experiments using UCI Phishing Websites Dataset. Performance is compared to the existing methods of artificial intelligence and machine learning and deep learning using a number of metrics, which indicate that the proposed method has greater accuracy, precision, recall, F1-score and AUC.

A. Dataset Description

This study uses the UCI Machine Learning Repository Phishing Websites Dataset which is a publicly available dataset that is a benchmark in the phishing detection research literature. The dataset consists of a total of 11,055 instances with 30 features that are used to identify characteristics of phishing and legitimate websites. The features include lexical features of URLs, content-based features identifying abnormal tags of the html content, and domain based features such as age of the URL and SSL certificate status. Each instance gets marked as phishing or legitimate, and it is possible to take supervised learning approaches. The dataset is very amenable to hybrid feature extraction and classification as it contains both structural and domain based behavioral features. In addition, both phishing and genuine website instances are equally represented in order to set up an unbiased foundation for the evaluation of the classification performance. The multifaceted features allow for robustness in experiments of hybrid detection approaches and thus the dataset could serve as standard benchmark for comparison of new phishing detection framework evaluation against existing state of the art approaches.

B. Performance Evaluation

The effectiveness of the proposed hybrid feature extraction and classification framework has been evaluated and it was compared against several state-of-the-art methods including traditional machine learning models and advanced deep learning methods. Comprehensive evaluation metrics of Accuracy, Precision, Recall, F1-Score and Area under the Curve (AUC) were used to compare the methods. The results are shown in Table 1 which shows the superior performance of the proposed method in comparison with the baseline models in terms of the phishing detection and robustness against the phishing attacks.

Table 1 Performance Analysis of the proposed method

Method	Accuracy (%)	Precision	Recall	F1-Score	AUC
Proposed Hybrid Method	98.2	0.981	0.983	0.982	0.992
Random Forest (RF)[14]	95.6	0.949	0.953	0.951	0.970
Support Vector Machine (SVM)[15]	94.3	0.940	0.941	0.940	0.965
Gradient Boosting (XGBoost) [16]	96.8	0.963	0.967	0.965	0.981
Deep Neural Network (DNN)[17]	97.1	0.971	0.969	0.970	0.985
CNN-LSTM Hybrid Model[18]	97.6	0.974	0.976	0.975	0.987

The results shown in the table 1 shows that the Proposed Hybrid Method have highest functionality out of all the compared methods with an accuracy of 98.2% and an AUC of 0.992- Indicative of excellent separability. Among the baseline models, CNN-LSTM model has the highest accuracy (97.6%), followed by Deep Neural Network (97.1%) and Gradient Boosting (96.8%). In comparison, traditional machine learning methods, Random Forest (95.6%) and Support Vector Machine (94.3%) have a lower performance than the previous models, and still considered suitable for phishing detection.

The evaluation using precision, Recall, and F1-scores showed the better functionality of deep learning and hybrid model than classical models to resolve phishing detection. Overall, the results suggest that the combination of lexical, content based, domain and behavioral features used with ensemble stacking classifier is more effective in detecting phishing than using the individual model approach only and provides additional robustness and flexibility.

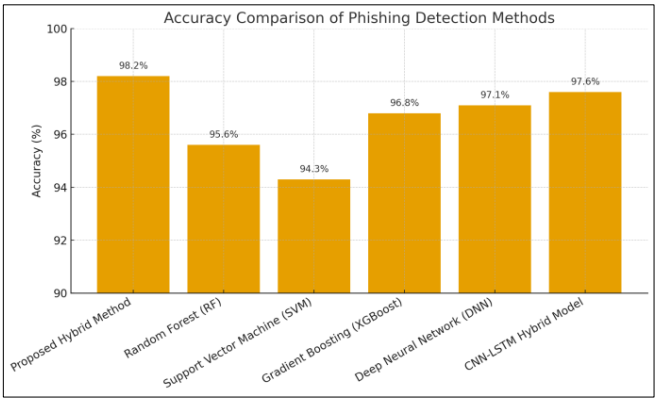

Figure 3 Performance Analysis (Accuracy) –Phishing Detection Methods

Figure 3 show that the proposed Hybrid Method is the highest accuracy phishing detection method by an accuracy percentage of 98.2%. The Proposed Hybrid Method is more accurate than all the alternatives methods presented on the bar chart, including CNN-LSTM Hybrid Model (97.6%) and a DNN (97.1%), which is the second and third highest accuracy method. The accuracy of the traditional machine learning approaches such as RF and SVM is noticeably less accurate that the accuracy of the proposed approach (RF) and SVM is 95.6% and 94.3% respectively. It is evident from the chart, using a deep learning framework with multiple models is more effective for phishing detection than the single algorithm approach.

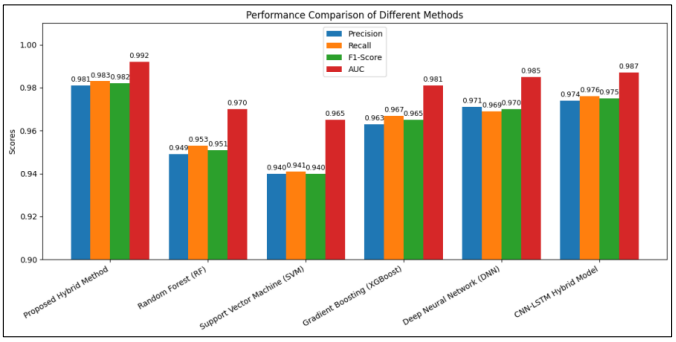

Figure 4 Performance Analysis(Precision, Accuracy, F1-Score and AUC)

From figure 4 , it is clear to see that the hybrid Method presented outperforms all the models used for base line in all four metric categories, reaching near perfect values for Precision, i.e. 0.981, Recall i.e. 0.983, F1-Score i.e. 0.982 and AUC i.e. 0.992. This has suggested not only an accurate classifier, but one which demonstrates reliability in distinguishing phishing web sites from non-phishing web sites. Within the models that used as the baselines the CNN-LSTM Hybrid Model has the best performance (Precision: 0.974, Recall: 0.976, F1: 0.975, AUC: 0.987) followed by the DNN and Gradient Boosting (XGBoost) which both achieve good and balanced performance. On the other hand, performance from traditional ML models such as RF and SVM have lower values for all the metrics as SVM have the weakest performance (Precision: 0.940, AUC: 0.965).

Overall, we can conclude that both the hybrid approach and DL based models both provide a statistically significant detection accuracy and reliability advantage over traditional machine learning approaches.

C. Discussion and Limitations

The proposed hybrid feature extraction and classification scheme exhibits a good performance in the identification of phishing web sites, outperforming other currently-existing systems in terms of overall accuracy, recall, and robustness. By extracting and employing a combination of lexical, content-based, domain-level and behavioral features, the proposed architecture is able to capture the various techniques used by attackers. In addition, the use of ensemble stacking classifier leads to improvement in adaptability from overfitting, and is a reliable detection framework in different phishing situations for general use. The experiments, in conjunction with a lot of other work, prove that it will accurately identify new attempts of phishing, and can easily be used for real-world implementations such as browser plug-ins, e-mail filters, and enterprise systems and applications.

Whilst the results have been promising, there have been limitations. First, the model is trained on publicly available datasets such as the UCI Phishing Websites Dataset which may not be representative of the evolving phishing strategies, most importantly, the highly sophisticated attacks which use AI-generated content. Second, in order to deploy the framework in the real world, retraining would be needed continuously in the event that the attack vectors evolve rapidly, which would add an additional computational overhead. Third, the framework requires stable access to information on domain registration and SSL certificates, which is usually not available in practice. Lastly, if the user selects one of the ensemble models the misclassification performance (acceptable level of accuracy) is high however they are usually less interpretable, which would be problematic as cyber security analysts need to have some means to explain their decisions. These need to be sorted out for practical large-scale application.

V. CONCLUSION

This study was a hybrid approach of feature extraction and classification to identify phishing websites, which combined lexical features, content features, domain features and behavior-based features, to provide a higher comprehensive representation of web threats. Using ensemble stacking classifier, we achieved the best performance with a comparison to baseline methods, if being able to deliver the accuracy and robustness to different phishing approaches. These results show that by combining the use of unconventional features with an ensemble learning methodology, any web threat detection system can achieve increased reliability, adaptability and real-world applicability to cybersecurity routines. Nevertheless, in as much as a change in phishing strategies are constant; continuing with the use of static datasets and handcrafted features cannot suffice. Future efforts and next steps on our research should focus on deep learning model cracking semantic and contextual features from

raw website data in a timely manner, real-time webpage screenshots, screen capture parsing, and even generative AI content. Real-time adaptive learning mechanism and reinforcement learning mechanism can be combined with the feedback loop to help the models dynamically update in case of new attack vectors. In addition, efforts should be made to incorporate explainable AI (XAI) to increase the interpretability, communication and trustworthiness to cybersecurity professionals penalized on trust. Finally, studies in large-scale deployments among different platforms and user environments are critical to reporting of scalability, generalizability, and long-term effectiveness of phishing detection framework designs.

REFERENCES

[1]. Ansar, Syed Anas, Jaya Yadav, Sujit Kumar Dwivedi, Ankur Pandey, Savarni Prakash Srivastava, Mohammad Ishrat, Mohd Waris Khan, Dhirendra Pandey, and Raees Ahmad Khan. "A critical analysis of fraud cases on the Internet." *Turkish Journal of Computer and Mathematics Education* 12, no. 12 (2021): 2164-2186.

[2]. Dutta, Ashit Kumar. "Detecting phishing websites using machine learning technique." *PloS one* 16, no. 10 (2021): e0258361.

[3]. Zaimi, Rania, Mohamed Hafidi, and Mahnane Lamia. "A deep learning approach to detect phishing websites using CNN for privacy protection." *Intelligent Decision Technologies* 17, no. 3 (2023): 713-728.

[4]. Alshingiti, Zainab, Rabeah Alaqel, Jalal Al-Muhtadi, Qazi Emad Ul Haq, Kashif Saleem, and Muhammad Hamza Faheem. "A deep learning-based phishing detection system using CNN, LSTM, and LSTM-CNN." *Electronics* 12, no. 1 (2023): 232.

[5]. Tang, Lizhen, and Qusay H. Mahmoud. "A deep learning-based framework for phishing website detection." *IEEe Access* 10 (2021): 1509-1521.

[6]. Said, Yahia, Ahmed A. Alsheikhy, Husam Lahza, and Tawfeeq Shawly. "Detecting phishing websites through improving convolutional neural networks with Self-Attention mechanism." *Ain Shams Engineering Journal* 15, no. 4 (2024): 102643.

[7]. Kalabarige, Lakshmana Rao, Routhu Srinivasa Rao, Ajith Abraham, and Lubna Abdelkareim Gabralla. "Multilayer stacked ensemble learning model to detect phishing websites." *Ieee Access* 10 (2022): 79543-79552.

[8]. Moedjahedy, Jimmy, Arief Setyanto, Fawaz Khaled Alarfaj, and Mohammed Alreshoodi. "CCrFS: combine correlation features selection for detecting phishing websites using machine learning." *Future Internet* 14, no. 8 (2022): 229.

[9]. Almomani, Ammar, Mohammad Alauthman, Mohd Taib Shatnawi, Mohammed Alweshah, Ayat Alrosan, Waleed Alomoush, and Brij B. Gupta. "Phishing website detection with semantic features based on machine learning classifiers: a comparative study." International Journal on Semantic Web and Information Systems (IJSWIS) 18, no. 1 (2022): 1-24.

[10]. Rao, Routhu Srinivasa, Amey Umarekar, and Alwyn Roshan Pais. "Application of word embedding and machine learning in detecting phishing websites." Telecommunication Systems 79, no. 1 (2022): 33-45.

[11]. Das Guptta, Sumitra, Khandaker Tayef Shahriar, Hamed Alqahtani, Dheyaaldin Alsalman, and Iqbal H. Sarker. "Modeling hybrid feature-based phishing websites detection using machine learning techniques." Annals of Data Science 11, no. 1 (2024): 217-242.

[12]. Kasim, Ömer. "Automatic detection of phishing pages with event-based request processing, deep-hybrid feature extraction and light gradient boosted machine model." Telecommunication Systems 78, no. 1 (2021): 103-115.

[13]. Wen, Tingke, Yuanxing Xiao, Anqi Wang, and Haizhou Wang. "A novel hybrid feature fusion model for detecting phishing scam on Ethereum using deep neural network." Expert Systems with Applications 211 (2023): 118463.

[14]. Yang, Rundong, Kangfeng Zheng, Bin Wu, Chunhua Wu, and Xiujuan Wang. "Phishing website detection based on deep convolutional neural network and random forest ensemble learning." Sensors 21, no. 24 (2021): 8281.

[15]. Anupam, Sagnik, and Arpan Kumar Kar. "Phishing website detection using support vector machines and nature-inspired optimization algorithms." Telecommunication Systems 76, no. 1 (2021): 17-32.

[16]. Gohil, Nayanaba Pravinsinh, and Arvind D. Meniya. "Click ad fraud detection using XGBoost gradient boosting algorithm." In International Conference on Computing Science, Communication and Security, pp. 67-81. Cham: Springer International Publishing, 2021.

[17]. Anitha, J., and M. Kalaiarasu. "A new hybrid deep learning-based phishing detection system using MCS-DNN classifier." Neural Computing and Applications 34, no. 8 (2022): 5867-5882.

[18]. Butt, Husnain Mansoor, Hasaan Haider, Marium Mehmood, and M. Asad Nadeem. "A Detecting Phishing URLs using LSTM-CNN hybrid Deep Learning Model." International Journal for Electronic Crime Investigation 9, no. 1 (2025).