SCIENTIFIC CULTURE, Vol. 12, No.5.1, (2026), pp. 350-359
Open Access. Online & Print

www.sci-cult.com

DOI: 10.5281/zenodo.12511030

EVALUATING PHISHING DETECTION SYSTEMS AGAINST OBFUSCATED URLS: A BENCHMARK AND ANALYSIS

Prof. Sunil Kumar Punjabi[1*]

[1*]*Research Scholar, Shridhar University, Pilani*
Assistant Professor, Dept. of Computer Engineering,
SIES Graduate School of Technology, Nerul, Navi Mumbai.
Mail ID: sunilp@sies.edu.in

ABSTRACT

Cyber security is an arms race between the bad guys and phishing protection systems, as the former continue to seek ways to obfuscate URLs. Machine learning techniques are highly effective in detecting phishing URLs, but it remains to be seen how they will hold up under attack. In this paper we report baseline testing of phishing protection systems for URL obfuscation. We developed a quantitative set of experiments using a list of URL features from the PhiUSIIL data set. We tested four machine learning models (Logistic Regression, Random Forest, Extreme Gradient Boosting (XGBoost) and Multi-Layer Perceptron (MLP) using baseline test data (a mix of good and bad URLs) and synthetic test data (a mix of good and bad URLs with medium obfuscation, including URLs with subdomain injection, character replacement and encoding). We found all models performed extremely well on baseline (better than macro F1 score of 0.99), but less well on obfuscated. We found Random Forest and XGBoost models were more robust, attaining more than 40% of original performance, while Logistic Regression and MLP were less robust (less than 20%). This is a trade-off between performance and robustness. This work shows the importance of robust models for phishing detection systems and provides a method for evaluating model robustness to inform the development of robust models for detecting phishing under attacks.

Keywords: *phishing detection, URL obfuscation, adversarial robustness, machine learning, character-level features, URL semantics, cybersecurity.*

1. INTRODUCTION

Phishing is one of the most widespread and active cybersecurity risks that exploit false URLs and web interfaces to acquire access to confidential user information. Since the online channels have increased, so have the effectiveness of the hackers in coming up with better malicious URLs that are quite similar to legitimate ones and in most instances, the difference in structure and lexical variations are not very high to be detected. The traditional techniques such as blacklist-based filtering and rule-based systems have struggled to keep up with the scale and adaptability of the present phishing applications. Thus, machine learning and deep learning techniques have been applied widely to increase the quality of detection and scalability. Recent studies indicate that machine learning models are capable of learning discriminative patterns, using URL-based features, and can be used rather effectively to perform phishing detection tasks [1]. At the same time, advances in artificial intelligence have led to the exploration of large language models to detect phishing, whereby systems can obtain contextual- and semantic-based information not limited to the traditional feature engineering frameworks [2]. Dataset-based frameworks, such as PhiUSIIL, are also helpful in this development, as they provide structured features representations and incremental learning capabilities allowing them to adapt to the dynamic threat conditions [3]. Phishing attacks are also dynamic, which explains the development of the adaptive learning strategies. Continuous learning techniques aim to

address concept drift by allowing models to acquire new knowledge without re-training them completely so that they remain useful over time [4]. On the same note, it has suggested that real-time detection systems should be used to efficiently process web traffic at large scale and offer timely detection of malicious URLs [5]. The foregoing developments highlight the increased significance of not only accurate but also responsive as well as detection systems, in line with the constantly evolving patterns of attacks.

Despite such developments, there has been an issue of phishing being detected, as there exists the adversarial behaviour of the attackers. Domain detection methods based on machine learning have been shown to detect suspicious domains, but their performance usually relies on rather stationary feature distributions [6]. Based on the comprehensive literature reviews, most of the existing approaches are based on improving the quality of classification, but a comparative lack of emphasis is put on the measurement of the resilience to the adversarial manipulation [7]. The fact that large scale phishing detection systems have been previously done has shown that it is possible to detect phishing in real time, but has also shown that it is very difficult to cope with the ever evolving threats [8]. More recent literature has begun to directly address the idea of evasion strategies, and models to compute the same in a manner that is not useless in adversarial contexts have been introduced [9]. Practical tests, in particular those involving log-in based phishing tests, also indicate that the model performance can significantly change under realistic conditions of attack as compared to controlled datasets [10].Though machine learning-based phishing detection system is almost perfect when the system is evaluated in a typical benchmark setting, their strength in an adversarial setting is underestimated. Practically, the attackers will carefully manipulate the URLs, to evade the detection systems by maliciously changing the URL structure by using character substitution, subdomain injection and encoding techniques. This creates a discrepancy between benchmark evaluation and real-world effectiveness since educated models may not be effective with obfuscated inputs. Moreover, other methods are based on webpage-level or content-based characteristics, which can be inaccessible at the early stage of detection, causing performance estimates to be not representative of the real implementation circumstances. Strongness to evasion techniques have been beginning to be explored in the literature; systematic benchmarking plans to evaluate model behaviour in the context of controlled obfuscation have been few, and not systematic schemes to compare and evaluate model behaviour in the context of controlled obfuscation.

In order to address such issues, the paper targets phishing detection on the basis of URL-based features, and it should be mentioned that the detection will be performed at the first levels, and the detection will be performed without the references to the webpage content or user interaction details. Structured numerical representations of URLs, such as lexical, structural, and obfuscation-related features are used to conduct the analysis. To minimize the leakage of features, the study will make use of the URL-level information to restrict the space of features and provide a more realistic approximation of the detection capabilities. The experimental design implies applying different machine learning models which represent different learning paradigms and compare the performance of these models under both the standard and adversarial conditions. Synthetic obfuscation is controlled and added to look like realistic alterations to URL structure, in order to perform a systematic test of the robustness of the models.

This research value is that it dispels the focus of the accuracy-based evaluation and shifts it to the robustness-based benchmarking. High accuracy may be applicable but it does not necessarily imply reliability in a court setting. This paper provides a systematic approach to learning about the behaviour of machine learning models manipulated by URL attributes by proposing a controlled obfuscation system. One can suppose that the findings will reveal the variation in the robustness of the different kinds of models and the ensemble-based method can be more robust as compared to the linear and neural models. Better detection systems to counter phishing can be created using such teachings and help in creation of a more realistic evaluation system. The primary aims of this study which are reached using the experimental design and analysis are:

- Compare the original performance of different machine learning models on phishing URL using a structured and URL-based feature set.
- Construct a regulated obfuscation system to imitate adversarial adjustments in phishing URLs.
- Evaluate and compare the strength of the models that have been tested regarding their performance when subjected to obfuscated conditions.

2. LITERATURE REVIEW

Recent developments in phishing detection have also gone towards taking advantage of deep learning and hybrid machine learning to enhance

the accuracy and flexibility of detection. Temporal modelling techniques, including Temporal Convolutional Networks (TCNs), have also been studied to establish sequential regularities in the URL structure and have been shown to model temporal dependencies and enhance classification accuracy in tasks of phishing URL detection [11]. Equally, it has been suggested that hybrid architectures that combine deep neural networks and recurrent models (like DNNLSTM models) can be used to achieve better feature representation through the combination of both spatial and sequential learning functions, which increases the accuracy of detecting phishing in a wide range of phishing samples [12].

Simultaneously, the conventional machine learning approaches still have a considerable role in detecting phishing. It has been observed that feature sets with care or delicate programming along with classical algorithms are capable of high detection accuracy especially when optimistic feature selection approaches are utilized [13]. Further, fusion of phishing detection into larger intrusion detection tools has been explored and it has been noted that integrated security systems have the potential to handle a multitude of cyber threats concurrently [14]. Phishing detection systems based on machine learning have also highlighted the significance of the feature variety and model optimization and proven that a mixture of multiple features categories can be used to achieve a stronger classification [15].

The deep learning-based methods have also enhanced the functions of the phishing detectors by allowing automatic generation of features out of raw data. More recent methods into this field have examined how deep neural networks can be used to recognize intricate patterns in phishing websites, with better results than more traditional techniques do [16]. In addition, real-time detection models, such as models of recurrent neural networks and character-level embeddings, have been presented to deal with streaming data, which is useful in dynamical environments to quickly detect malicious URLs [17]. Optimization techniques to train deep learning models have also been improved to enhance performance as research continues to enhance the model architecture and training strategies to enhance accuracy of detection and generalization [18].

Besides better performance, interpretability and explainability have also been introduced to the fore in phishing detection research. The efforts to visualize and analyse the decision-making of the recurrent neural networks have served to understand the way the model identifies phishing patterns thereby increasing the level of transparency and trust in automated systems [19]. Relaxed character and word-level modelling approaches have been suggested to add intuitively motivated lexical schemes to URLs and demonstrate how deep learning can be successfully applied to process complex and obfuscated inputs [20].

Some of these advances were predetermined by previous background research that showed the viability of machine learning-based malicious URL detection with large-scale data sets and feature-based methods [21]. These studies described the importance of lexical and host-based features to make classification, which continues to define the current phishing-detection techniques. This has been extended later by studies that have incorporated features of the hyperlink to further improvements of feature engineering strategies to enable machine learning models to be more efficient in detecting phishing websites [22].

All in all, the literature has shown a definite advancement of the simple feature-based machine learning methods to advanced deep learning and hybrid models that are able to capture complex variations in phishing URLs. Despite the fact that the existing methods have achieved high accuracy rate and enhanced flexibility, the problem of ensuring that the methods are not subject to adversarial manipulations such as URL obfuscation still exists. This drives the need to have systematic evaluation frameworks that is not only able to assess classification performance, but also to examine under realistic attack conditions the resilience of the model.

3. METHODOLOGY

3.1 Research Design

The paper has employed quantitative experimental benchmarking study in comparing the phishing detection systems with the obfuscated URLs. It was a comparative design since several machine learning models were tested on the same experiment and diagnostic since the test was carried out to determine the performance degradation and failure on robustness.

The set up of the experiment was in two conditions of assessment:

- Base line constituting: initial performance mixed test set.
- Stress-test: setting Performance on a synthetically obfuscated test set.

This configuration made it possible to compare directly normal performance on detected data with robustness on adversarial manipulation of URLs.

3.1.1 Benchmarking Objective

The methodological goal was not only to measure predictive accuracy, but to test the predictive performance of the phishing detection models

under the circumstances that the URL structures are intentionally deformed to look like the realistic obfuscation techniques that are used in phishing attacks.

3.1.2 Low-Leakage Evaluation Principle

In order to sustain the benchmark at the same level with the first stage of phishing detection, the experiment suggested the low-leakage design. End modelling only maintained URL-based properties and dropped the webpage-content or post-visit property to avoid the proliferation of unrealistic cues.

3.2 Data Source and Collection Method

The PhiUSIIL Phishing URL Dataset [23], which has labelled URLs and constructed features based on URL structure, domain composition, and obfuscation, was used to conduct the study. The information was taken by a preexisting source of benchmarks; hence, the data gathering procedure was not field based but secondary.

3.2.1 Data Validation

Prior to experimentation, the dataset was analysed in terms of duplicate records, missing values, feature types, label consistency and structural integrity on a whole. Also, reconsideration of the label semantics was conducted before developing the model to make sure that phishing and legitimate classes were interpreted correctly.

3.2.2 Feature Selection

The final benchmark used a URL-centred subset in spite of the fact that many engineered variables are in the dataset but had to remain within the scope of the title of the paper. The features that were chosen were:

- URL length
- domain length
- number of subdomains
- obfuscation presence
- obfuscation ratio
- digit-related measures
- special-character measures
- HTTPS usage

This feature filtering made sure that the evaluation was focused on URL-based phishing detection and not on page-content analysis.

3.3 Population and Sampling

The target population of the study were the URLs that can be classified as phishing or legitimate in a cybersecurity detection scenario. Since the study utilized a benchmark dataset, the population of analysis was restricted to the records that are structured by the URL that are found in the PhiUSIIL dataset. The stratified train-test split was used to maintain the proportion of classes in data and 80 percent of the data was used during training and 20 percent during testing. The stratification was required due to the frequent class imbalance of phishing datasets, and preserving the initial class distribution allows having a fair and representative assessment of model performance. This was followed by the setting up of two test environments; the original mixed test set was used to evaluate baseline evaluation and the synthetic test set of stress-testing evaluation. The stress-test enrolment was made out of legitimate test sample and synthetically obfuscated phishing sample in manner that the resulting test could be considered as a valid binary classification problem.

3.4 Controlled Obfuscation Generation

In order to test the resilience against masquerised URLs, another controlled synthetic obfuscation step was added. The test set phishing URLs were obfuscated using a set of obfuscation techniques that were selected to create realistic attacker behaviour.

3.4.1 Obfuscation Techniques

The obfuscation phase was controlled and used four transformation strategies on phishing URLs, including percent encoding, subdomain injection, character substitution, and noise-token injection. These methods have been chosen in order to replicate realistic URL obfuscation patterns that are used in phishing attacks to hide malicious links and avoid detection mechanisms.

3.4.2 Feature-Space Perturbation

Since the predictive models worked on structured tabular features instead of raw URL strings, the process of obfuscated appeared in feature space by changing the values of variables, including URL length, domain length, subdomain count, obfuscated character count, obfuscation ratio, number of digits and number of special-characters ratio. This provided regularity between the resulting pattern of obfuscation and the feature description applied in the evaluation of models.

3.5 Machine Learning Models

There were four trained classification models that were used to describe the different learning paradigm: logistic regression to represent a linear classifier as a baseline, Random Forest as an ensemble classifier in the form of a bagging algorithm, XGBoost as an ensemble classifier based on boosting, and Multi-Layer Perceptron (MLP) as a neural network classifier. Such a model choice allowed making a direct comparison between linear, ensemble, and neural methods in a similar framework of benchmarking.

3.5.1 Data Preparation

Numeric values were also handled with median imputation in order to deal with missing data prior to the training, and where it was important, standardization ensured to both the Logistic Regression and MLP models. This preprocessing provided uniform and equitable training of the model with all classifiers, which provides an opportunity to easily compare their performance.

3.6 Data Analysis Technique

The original mixed test set was initially assessed in terms of the accuracy, balanced accuracy, macro F1-score, phishing precision, phishing recall, and phishing F1-score. The confusion matrices were also investigated to learn the behaviour of class-level prediction. The same trained models were then tested on the synthetic stress-test set with the same metrics to be able to compare directly standard conditions and obfuscated conditions.

3.6.1 Robustness Metrics

To quantify resilience under obfuscation, two robustness indicators were calculated:

- Contact F1 Drop = baseline macro F1 - stress-test macro F1.
- F1 Robustness Ratio = micro F1/macro F1 stress-test macro F1 baseline macro F1. These are those measures that ranked models based on robustness.

3.6.2 Comparative Interpretation

The analysis was aimed at identifying the model that maintained the highest performance under obfuscation, models that had the worst performance degradation, and the extent of robustness across the model families. Also, tree-based models were analyzed using feature importance and the attributes of the URLs that most strongly contributed classification performance were identified.

3.7 Ethical Considerations

This research was not done with human subjects, personal information gathering, interviews, and face-to-face interaction. Nonetheless, ethical considerations were also noteworthy since the study deals with phishing behaviour and manipulations with adversarial URLs.

The dataset has been employed in purely academic and defensive purposes of cybersecurity studies. Synthetic obfuscation step was developed to obstruct only the offline model testing and not to produce deployable or harmful links to phishing. Each and every experiment was performed under controlled conditions, and the research was also reported clearly to not overestimate the power of models and promote their abuse.

4. RESULTS

4.1 Baseline Performance on Original Mixed Test Set

The original mixed test set base evaluation shows that all four models performed almost perfectly in the most common measures. Random Forest and XGBoost scored high on macro F1, accuracy, balanced accuracy, phishing precision, phishing recall and ROC-AUC which means that phishing and non-phishing URLs were perfectly separated under normal circumstances. Logistic Regression recorded a macro F1 of 0.995903 with MLP recording 0.997899 with small misclassifications.

These findings are supported by the confusion matrices, which reveal that the linear and neural models only have low false positives or false negatives. These findings suggest that the chosen URL-based feature set is quite informative and can be used to learn the boundaries of discriminative decisions. Table 1 presents the baseline performance indicators, which are used as a benchmark in the further robustness analysis.

Table 1. Baseline and Moderate Obfuscation Performance Comparison

Model	Macro F1 (Original Mixed Test)	Macro F1 (Synthetic Stress Test)	F1 Drop	Robustness Ratio
XGBoost	0.996858	0.444968	0.551890	0.446371
Random Forest	0.997205	0.405313	0.591892	0.406449
Logistic Regression	0.995903	0.166615	0.829288	0.167301
MLP	0.997899	0.166662	0.831237	0.167012

4.2 Performance Under Moderate URL Obfuscation

The synthetic stress-test set was made so that it appears to represent realistic URL obfuscation patterns, such as subdomain injection, percent encoding, character substitution and token noise injection. Training of the trained models using this dataset found that there were significant variations

in resilience. Random Forest and XGBoost did not significantly drop macro F1 to 0.405313 and 0.444968 and Logistic Regression and MLP dropped to 0.166615 and 0.166662.

This disparate performance brings out the influence of the obfuscated models on paradigm of learning models. Both the baseline and stress-test measures are shown in Table 1, as well as the F1 drop and the robustness ratio, whereas the extent of the F1-score decrease among the models is visually illustrated in Fig. 1. These unanimous results indicate that ensemble-based methods are resistant to feature interference, when compared to the linear and neural models, which are extremely vulnerable to obfuscation.

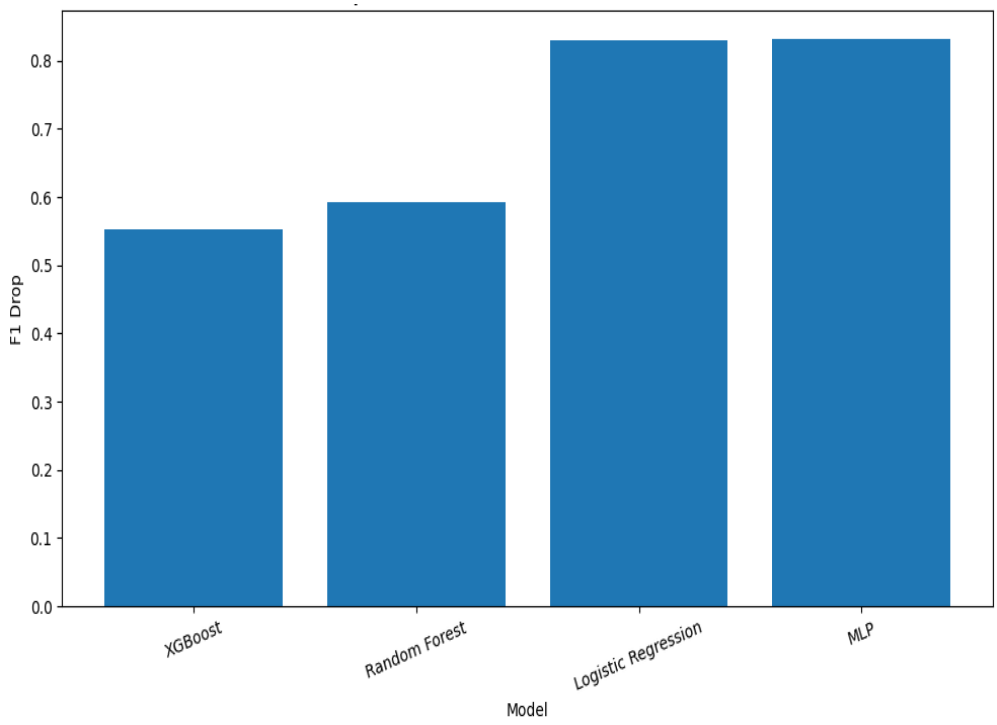

Fig. 1. Summary of Robustness Under Moderate URL Obfuscation.

4.3 Comparative Robustness Across Models

Other comparisons among models can also be appreciated by the robustness analysis presented in Table 2 and Fig. 2. XGBoost and Random Forest had the lowest F1 drops (0.551890 and 0.591892, respectively) and the best ratio of robustness (0.446371 and 0.406449), which means that they are the most capable to face shifts in features caused by obfuscation. The significantly higher F1 decreases (0.829288 and 0.831237) and extremely low robustness ratios (approximately 0.167) of the Logistic Regression and MLP indicated their susceptibility to adversarial examples.

This dichotomy of two types shows that model architecture is an important criterion of robustness. The ensemble approaches that are based on non-linear boundaries of decision making are effective in absorbing the variations in the URL feature distributions; the linear and shallow neural networks do not work in the situation characterized by adjustments.

These patterns can be demonstrated by the introduction of tables and figures into the text. As an example, Fig. 1 is a comparative graphic sketch of the drops in F1-score that reveal the drops of the Logistic Regression and MLP are steep compared to those of XGBoost and random forest. Fig. 2 presents the ratio of robustness and it is possible to compare the retained baseline performance across models. Such visualizations complement the table results and enable one to learn intuitively the model behaviour under stressful situation. The aggregation of the analysis confirms that the analysis of the operational readiness of the phishing detection systems needs the robustness-oriented assessments as the general measures of the original data set may create an illusion of the overall reliability.

Table 2. Model Ranking Based on Robustness Under Obfuscation

Rank	Model	F1 Drop	Robustness Ratio	Interpretation
1	XGBoost	0.551890	0.446371	Most robust
2	Random Forest	0.591892	0.406449	Strong robustness

3	Logistic Regression	0.829288	0.167301	Weak robustness
4	MLP	0.831237	0.167012	Weakest robustness

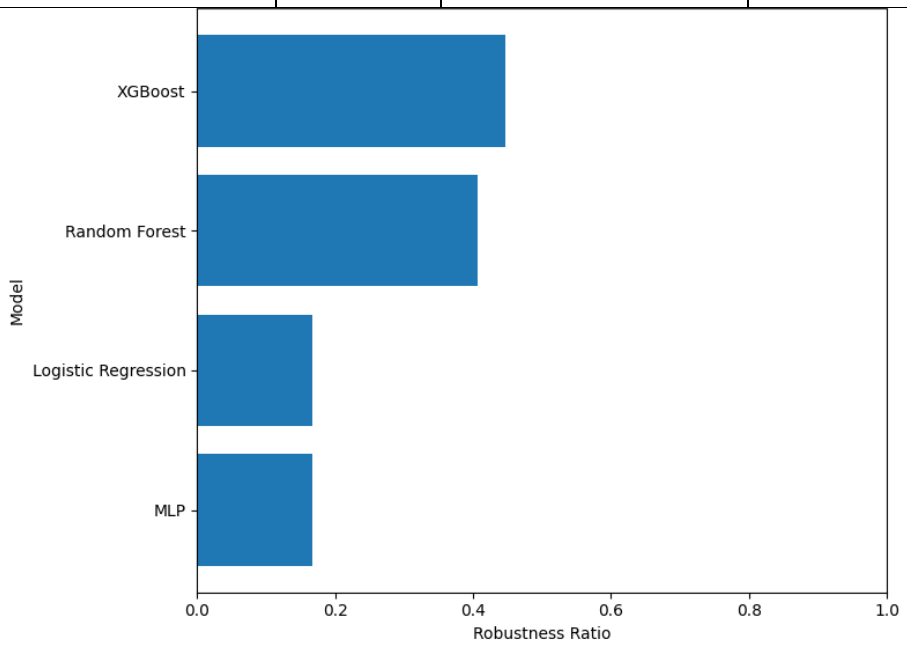

Fig. 2. Model Robustness Ratio Under Moderate URL Obfuscation.

In all models, the tendencies show that the baseline performance is exceptionally high, but the vulnerability to URL obfuscation differs greatly. The ensemble based models are always more robust than the linear and neural models, indicating that they are more appropriate in the adversarial environment. Besides, the stress-test outcomes prove that it is not enough to consider detection systems on the basis of accuracy or F1-score only to assess their effectiveness in standard conditions. The resilience of a phishing detection system can only be comprehended through the addition of controlled obfuscation and the study of the performance degradation.

On the whole, the findings affirm the purpose of the study of benchmarking phishing detection systems against obfuscated URLs and offer practical information on model selection and design. The combined representation of Table 1, Table 2 and Figs. 1 and 2 make sure that quantitative performance is presented as well as comparative robustness, which is vital to the research of cybersecurity.

5. DISCUSSION

This finding of this research is a great disparity between the base performance and the resilience in adverse situations. Though the tested models were able to achieve a score of virtually-perfect macro F1-scores on the mixed original test set, they dropped down to extremely-lowe-levels of performance when presented with synthetic obfuscation. It implies that the high accuracy on the standard benchmark environments does not reflect the reliability in adversarial manipulation. This degradation implies that the models are very sensitive to the presence of stable features and distribution of the same features which is distorted in case the phishing URLs are obfuscated. The tested models that were found to be more robust were XGBoost and Random Forest models, where the clearing percentage of higher percentage of, baseline performance was found to be made. This is because ensemble-based methods can provide non-linear and finer relationships and respond more to varies inputs attributes. On the other hand, MLP and Logistic Regression were performing too poorly, and this was an indication of not being flexible to feature perturbation in the obfuscated features.

These results can be correlated with the previous studies, which provide the information about high-performance of machine learning and deep learning models in the phishing detection processes that are controlled. List of approaches based on temporal convolutional networks have demonstrated functionality in the detection of sequential arrangement in URL formats which contribute to higher classification rates [11]. On the same note, the use of hybrid systems that use deep neural networks coupled with recurrent elements has also been suggested to achieve an improved representation of features and model performance [12]. Phishing detection tasks based on traditional machine learning using well-engineered feature sets have also been identified to achieve high accuracy with feature selection optimization [13]. Moreover, intelligent systems of detection that are

based on machine learning have highlighted the value of having various categories of features to enhance predictive accuracy [15]. Through still more recent research, deeper learning-based methods have been investigated, demonstrating that these systems can be used to automatically identify complex patterns and enhance detection in phishing [16].

Nevertheless, these studies may point to the classification accuracy improvements, but, in general, they fail to discuss conditions of robustness in adversarial situations. The current research contributes to this literature by proving that even the higher working models become weak when presented to obfuscated inputs. The large decrease in performance of the models under the circumstances of stress implies that the acquired boundaries of decision-making are not material enough to include distinctions in the pattern of distribution. This is particularly weak with Logistic Regression and MLP models, which were still significantly degraded and only a small fraction of their original performance. These results confirm the hypothesis that the accuracy is no longer a sufficient indicator of the model efficacy in practice in cybersecurity. These results have significant implications but both in studies and with practical implementation. First of all, they point to the need of framework-based robustness-sensitive detection of phishing attacks. The models that work well in a static environment might not work when used in the environment where adversarial manipulation is prevalent. Second, it means that the results suggest that resilience to perturbation of features should be taken into account when choosing a model and predictive accuracy. Ensemble based systems like the XGBoost, random forest, seem to provide a better option when making a deployment based on their relative robustness. Third, the study demonstrates the importance of adversarial testing in the development of the model. Model evaluation under controlled conditions of obfuscation provides a more realistic perspective on the performance of the model and can be used to identify the weaknesses before the model is deployed since the conditions are controlled.

Notwithstanding its contributions, this study is limited in a number of ways. It is analyzed with the help of one dataset, and thus may not be detailed enough to reflect the different forms of phishing attacks that have been encountered in real life situations. The dataset is a structured and exhaustive set of features, but it is possible that its distribution will impact the model performance and limit the generalizability. Also, although the process of synthetic obfuscation is systematic and regulated, it might only be able to capture the complexity of real-world techniques of adversarial. The analysis also cannot use raw URL text or webpage content which can also contain other context to be identified and are structured and tabular. In addition, few standard machine learning models are tested and not regarded of more advanced models that could offer better robustness.

The future research can be improved by including additional datasets to increase the generalizability and represent a more exhaustive set of phishing attack patterns. The other way of significance is the development of more complex ways of adversarial simulating, which may be more precise in comparison with the ways of obfuscation in life. In addition, the integration of the state-of-the-art deep learning frameworks and more intense features can also enhance the capabilities of detection systems to detect different complex patterns and strengthen them. Some standardized benchmarking models, including adversarial evaluation, would also help to make more meaningful comparisons between studies. Overall, this paper can potentially prove the importance of applying methods outside of the traditional metrics of accuracy in evaluating phishing detection systems and the need to adopt rigor cognitive methods to cybersecurity research.

6. CONCLUSION

The research question being tested in the current study was the degree of performance of machine learning based phishing detection systems when the conditions of URL obfuscation are under control. Although the operation of all the experimented models was almost optimal in the utilization of the original mixed test set, the operation indicated a massive deterioration of functionality to the exposure of the obfuscated phishing URLs. The ensemble-based models and especially the XGBoost and the Random Forest remained comparatively high in the tier of the performance, and the performance was considerably diminished in the other models which include the Logistic Regression and MLP. These findings prove that high baseline accuracy does not necessarily relate to real world reliability especially when the opponents are eager to thoughtfully construct URLs so that they can get away with a game of high stakes. There are high implications of the research both to research and realistic implementation. It highlights the need to go beyond the accuracy-focused appraisal and embark on strength-conscious benchmarking. The detection systems used on the real world, must be capable of dealing with adverse perturbations, and thus, resiliency and predictive accuracy should be taken into consideration in the process of model selection.The success of the ensemble-based

models shown have shown that the increase of robustness in phishing detection can be obtained through the addition of non-linear style of learning to increase the model. Based on these results it is suggested that adversarial testing is a prescribed model evaluation system to be included in future phishing detection systems. Interestingly, the situations of controlled obfuscation must also be created during the development period to reveal the weaknesses and render the system more accountable. In addition, it has also been recommended that researchers should utilize standardized benchmarking schemes that feature significant metrics of robustness to ensure that they could have more meaningful comparisons between models and studies. Future studies would enhance the robustness testing area by ensuring that there are varied datasets and more realistic adversarial methods that the research would not have addressed as it is that would be more reflective of the real-world phishing methods. More detection can be achieved by incorporating even more sophisticated deep learning models and more expressive models, incorporating character incorporations and contextual incorporations. Overall, this paper proves the essentiality of the robustness in the phishing detection and provides the source to create more robust cybersecurity infrastructures.

REFERENCES

[1] H. Ghalechyan, E. Israyelyan, A. Arakelyan, G. Hovhannisyan, and A. Davtyan, "Phishing URL detection with neural networks: An empirical study," *Scientific Reports*, vol. 14, no. 1, p. 25134, 2024.
[2] T. Koide, H. Nakano, and D. Chiba, "ChatPhishDetector: Detecting phishing sites using large language models," *IEEE Access*, vol. 12, pp. 154381–154400, 2024.
[3] A. Prasad and S. Chandra, "PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning," *Computers & Security*, vol. 136, p. 103545, 2024.
[4] A. Ejaz, A. N. Mian, and S. Manzoor, "Life-long phishing attack detection using continual learning," *Scientific Reports*, vol. 13, no. 1, p. 11488, 2023.
[5] S. Asiri, Y. Xiao, S. Alzahrani, and T. Li, "PhishingRTDS: A real-time detection system for phishing attacks using a deep learning model," *Computers & Security*, vol. 141, p. 103843, 2024.
[6] S. Alnemari and M. Alshammari, "Detecting phishing domains using machine learning," *Applied Sciences*, vol. 13, no. 8, p. 4649, 2023.
[7] A. Safi and S. Singh, "A systematic literature review on phishing website detection techniques," *Journal of King Saud University – Computer and Information Sciences*, vol. 35, no. 2, pp. 590–611, 2023.
[8] S. Marchal, J. François, R. State, and T. Engel, "PhishStorm: Detecting phishing with streaming analytics," *IEEE Trans. Netw. Serv. Manage.*, vol. 11, no. 4, pp. 458–471, 2014.
[9] T. Kim, N. Park, J. Hong, and S. W. Kim, "Phishing URL detection: A network-based approach robust to evasion," in *Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS)*, 2022, pp. 1769–1782.
[10] M. Sánchez-Paniagua, E. F. Fernández, E. Alegre, W. Al-Nabki, and V. González-Castro, "Phishing URL detection: A real-case scenario through login URLs," *IEEE Access*, vol. 10, pp. 42949–42960, 2022.
[11] M. A. Remmide, F. Boumahdi, N. Boustia, C. L. Feknous, and R. Della, "Detection of phishing URLs using temporal convolutional network," *Procedia Computer Science*, vol. 212, pp. 74–82, 2022.
[12] A. Ozcan, C. Catal, E. Donmez, and B. Senturk, "A hybrid DNN–LSTM model for detecting phishing URLs," *Neural Comput. Appl.*, vol. 35, no. 7, pp. 4957–4973, 2023.
[13] M. Bahaghighat, M. Ghasemi, and F. Ozen, "A high-accuracy phishing website detection method based on machine learning," *J. Inf. Secur. Appl.*, vol. 77, p. 103553, 2023.
[14] R. Jayaraj, A. Pushpalatha, K. Sangeetha, T. Kamaleshwar, S. U. Shree, and D. Damodaran, "Intrusion detection based on phishing detection with machine learning," *Measurement: Sensors*, vol. 31, p. 101003, 2024.
[15] A. K. Jha, R. Muthalagu, and P. M. Pawar, "Intelligent phishing website detection using machine learning," *Multimedia Tools Appl.*, vol. 82, no. 19, pp. 29431–29456, 2023.
[16] U. Zara, K. Ayyub, H. U. Khan, A. Daud, T. Alsahfi, and S. G. Ahmad, "Phishing website detection using deep learning models," *IEEE Access*, vol. 12, pp. 167072–167087, 2024.
[17] M. A. Nuritdinovich *et al.*, "Real-time phishing URL detection using gated recurrent units and character-level embedding," in *Proc. Int. Conf. Next Generation Computing Systems (ICNGCS)*, 2025, pp. 1–6.
[18] K. Barik, S. Misra, and R. Mohan, "Web-based phishing URL detection model using deep learning optimization techniques," *Int. J. Data Sci. Anal.*, vol. 20, no. 5, pp. 4449–4471, 2025.
[19] T. Feng and C. Yue, "Visualizing and interpreting RNN models in URL-based phishing detection," in *Proc. ACM SACMAT*, 2020, pp. 13–24.
[20] F. Tajaddodianfar, J. W. Stokes, and A. Gururajan, "Texception: A character/word-level deep learning

model for phishing URL detection," in *Proc. IEEE ICASSP*, 2020, pp. 2857–2861.
[21] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Learning to detect malicious URLs," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–24, 2011.
[22] A. K. Jain and B. B. Gupta, "A machine learning based approach for phishing detection using hyperlinks information," *J. Ambient Intell. Humaniz. Comput.*, vol. 10, no. 5, pp. 2015–2028, 2019.
[23] A. Prasad and S. Chandra, "PhiUSIIL phishing URL dataset," *UCI Mach. Learn. Repository*, 2024. [Online]. Available: https://archive.ics.uci.ed u/dataset/967/phiusiil%2Bphishing%2Burl%2Bdataset