Adversarial Evasion of Ensemble Learning Models for Malicious URL Detection: A White-Box Analysis with FGSM Attacks

Roopesh Kumar B N [1], Rekha B Venkatapur [2], Suman B S [3], Anubhav Misra [4]
[1] Department of Computer Science & Engineering, K.S. Institute of Technology (affiliated to Visvesvaraya Technological University, Belagavi), Karnataka, India

[2] Department of Computer Science & Engineering, K.S. Institute of Technology (affiliated to Visvesvaraya Technological University, Belagavi), Karnataka, India

[3] Department of Computer Science & Engineering, K.S. Institute of Technology (affiliated to Visvesvaraya Technological University, Belagavi), Karnataka, India

[4] Department of Computer Science & Engineering, K.S. Institute of Technology (affiliated to Visvesvaraya Technological University, Belagavi), Karnataka, India

Abstract: *Machine learning-based malicious URL detection has become a cornerstone of modern cybersecurity, offering scalability and adaptability beyond traditional signature-based methods. Among ensemble models, Random Forest (RF) demonstrates strong classification performance, achieving 89.78% accuracy in detecting malicious URLs. However, its vulnerability to adversarial attacks raises significant concerns about its deployment in real-world threat landscapes. This study systematically analyses the robustness of RF under white-box adversarial attacks, specifically the Fast Gradient Sign Method (FGSM). These attack strategies manipulate URL features with subtle perturbations, leading to a measurable decline in detection accuracy. Empirical evaluations highlight critical weaknesses in RF-based detection systems, emphasizing the necessity for adversarial defense mechanisms. The findings reinforce the urgency of developing robust countermeasures to safeguard machine learning-driven cybersecurity solutions against evolving attack methodologies.*

Keywords: **Malicious URL Detection, Adversarial Machine Learning, Ensemble Learning, Random Forest, Gradient-Based Attacks, Fast Gradient Sign Method (FGSM), Cybersecurity, Model Robustness, White-Box Adversarial Attacks.**

1. Introduction

With the rapid expansion of the internet, websites have become an integral part of daily life, facilitating services such as social media, online banking, and e-commerce. These websites are accessed via Uniform Resource Locators (URLs), which serve as unique web identifiers. However, along with legitimate websites, a significant number of malicious URLs are actively deployed to deceive users into disclosing sensitive information or downloading malware [1]. Cybercriminals frequently exploit human vulnerabilities and poor security awareness, increasing the likelihood of cyberattacks. The growing sophistication of cyber threats has led to a surge in phishing attacks, malware distribution, and fraudulent schemes [2]. Reports indicate that in the first quarter of 2022, over one million phishing attacks were recorded, illustrating the scale and urgency of this issue [3]. Attackers often create spoofed websites that resemble legitimate entities, tricking unsuspecting users into entering sensitive credentials. As phishing and malware distribution techniques evolve, traditional security measures, such as blacklisting and signature-based detection, have become insufficient due to their inability to detect newly generated malicious URLs [4]. While heuristic-based approaches offer slight improvements by detecting anomalous URL structures, they remain ineffective against adversarial manipulations and zero-day attacks [5].

To overcome these limitations, machine learning (ML) and deep learning (DL) techniques have emerged as promising solutions for automated malicious URL detection [6].

ML-based methods analyze lexical features, network behavior, and webpage content to distinguish between benign and malicious links [7]. Unlike rule-based detection systems, ML models can identify patterns in large-scale datasets, improving detection accuracy and generalization to novel threats [8]. Additionally, deep learning architectures, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have demonstrated effectiveness in character-level analysis for phishing URL detection [9].

Despite these advancements, ML-based URL detection faces critical challenges such as class imbalance, generalization issues, and adversarial vulnerabilities [10]. One of the most pressing concerns is the emergence of adversarial attacks, where attackers introduce subtle perturbations to URLs to evade detection models [11]. Adversarial attacks like the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) subtly modify URL-based features to deceive ML classifiers [12]. Studies indicate that adversarial manipulations can reduce detection accuracy by more than 30%, severely impacting real-world deployment [13]. Unlike traditional threats, adversarial generated URLs retain their functionality while effectively bypassing ML-based security systems. This necessitates the development of robust countermeasures, including adversarial training, ensemble learning, and anomaly detection mechanisms [14]. This study focuses on a systematic evaluation of ML-based malicious URL detection models under FGSM adversarial attack [15].

Given the limitations of existing approaches and the evolving sophistication of adversarial attacks, this study aims to assess the robustness of machine learning-based malicious URL detection models in adversarial settings. Unlike previous studies that focus primarily on feature engineering and classifier optimization [16], our research systematically evaluates adversarial vulnerabilities in ensemble learning models, particularly Random Forest and XGBoost [17]. The results provide critical insights into the weaknesses of ML-based security systems, guiding future research in developing resilient adversarial defense mechanisms. By leveraging a comprehensive dataset and implementing controlled adversarial attacks, this study contributes to the broader domain of cybersecurity by enhancing the resilience of machine learning-driven malicious URL detection systems [18].

The implications of adversarial vulnerabilities in URL detection systems extend beyond isolated misclassifications. These evasion tactics can be weaponized by threat actors to bypass intrusion detection systems, distribute ransomware through phishing campaigns, or facilitate access to command-and-control servers. As such, the robustness of ML-based URL classifiers is not merely an academic concern but a critical factor in the resilience of broader cybersecurity infrastructures. This paper addresses this gap by systematically evaluating ensemble models under adversarial pressure to inform future deployment strategies.

2. Methodology

This study utilizes a publicly available dataset from Kaggle, comprising a diverse set of benign and malicious URLs to ensure a representative sample for training and evaluation. To enhance data quality and reliability, preprocessing steps were applied, including removal of duplicate entries, handling of missing values, and standardization of URLs to maintain format consistency. Given the inherent imbalance in dataset distribution, where benign URLs significantly outnumber malicious ones, up sampling techniques were employed to mitigate classification bias and enhance model generalization [12].

2.1 Feature Extraction and Preprocessing

Feature extraction was conducted by analyzing the lexical and structural properties of URLs, which have been widely recognized as discriminative features in malicious URL detection tasks. These features include URL length, the number of special characters, digit occurrences, entropy measures, and token-based segmentation [7]. After extraction, the features were normalized using StandardScaler, a widely adopted technique that ensures all features are on a comparable scale, thereby improving model convergence and classification performance [13].

2.2 Model Selection and Training

For classification, we employed Random Forest (RF) and XGBoost, two widely used ensemble learning algorithms known for their effectiveness in handling high-dimensional datasets and robustness in classification tasks [17]. The dataset was divided into 80% training and 20% testing subsets, following standard machine learning evaluation protocols. Hyperparameter tuning was conducted using grid search and cross-validation, optimizing parameters such as the number of estimators, maximum depth, and learning rate to maximize predictive performance.

2.3 Evaluation Metrics

The trained models were evaluated using standard performance metrics, including:

- Accuracy – Measures overall classification performance.
- Precision – Evaluates the proportion of correctly classified malicious URLs.
- Recall – Assesses the model's ability to detect malicious URLs effectively.
- F1-score – Provides a balance between precision and recall, particularly for imbalanced datasets [14].

These metrics offer a comprehensive assessment of the model's efficacy in distinguishing between benign and malicious URLs, ensuring its practical applicability in cybersecurity environments.

2.4 Adversarial Robustness Considerations

Ensemble models such as Random Forest and XGBoost are inherently resilient to minor perturbations due to their feature randomness and bagging mechanisms, making them suitable for malicious URL detection [15]. However, as adversarial threats evolve, these models remain susceptible to adversarial manipulations. This study systematically evaluates their vulnerabilities under gradient-based adversarial attacks, highlighting the need for advanced defense strategies to enhance model robustness.

Figure 1: Methodology

3. Dataset

The dataset used in this study was sourced from Kaggle and comprises a collection of benign and malicious URLs, ensuring a diverse and representative sample for training and evaluation. It contains a total of 549,346 URLs, categorized into two classes: malicious and benign. The dataset follows a highly imbalanced distribution, with 392,924 benign URLs (71.5%) and 156,422 malicious URLs (28.5%), reflecting real-world internet traffic trends where legitimate websites vastly outnumber harmful ones. Each entry in the dataset consists of two primary attributes: the URL itself and its corresponding class label (0 for benign and 1 for malicious). The benign URLs in the dataset originate from trusted sources, such as reputable websites and commonly accessed domains, while the malicious URLs are collected from cybersecurity threat intelligence platforms, phishing databases, and online blacklists. This ensures that the dataset includes a variety of malicious URLs associated with phishing, malware distribution, and fraud attempts.

Given the class imbalance, where benign URLs significantly outnumber malicious ones, resampling techniques such as oversampling of the minority class or synthetic data generation may be necessary to improve model generalization. The dataset is pre-processed by removing duplicates, handling missing values, and standardizing URL formats to enhance consistency and ensure reliable feature extraction. This dataset serves as the foundation for training and evaluating machine learning models for malicious URL detection, enabling the development of robust cybersecurity defenses against web-based threats.

Dataset	Source	Category	Malicious	Benign
Dataset1	Kaggle	Malicious, Benign	156,422	392,924

Table 1: Dataset Description

4. Result and Analysis

To evaluate the effectiveness of machine learning models in detecting malicious URLs, multiple classifiers were trained and tested on the original, unaltered dataset. The models were assessed using accuracy, precision, recall, and F1-score, which are widely used performance metrics in cybersecurity-related machine learning applications. The classifiers tested include Logistic Regression, SVM, KNN, Random Forest (RF), XGBoost, AdaBoost, and LightGBM. Among the traditional classifiers, Logistic Regression (80.44%) and KNN (81.88%) exhibited lower accuracy, indicating limitations in handling complex URL patterns.

The results demonstrate that ensemble learning models significantly outperform traditional classifiers. Random Forest (RF) achieved the highest accuracy (89.78%), followed by LightGBM (89.50%) and XGBoost (88.00%). These models leverage feature randomness, boosting mechanisms, and multiple decision trees, allowing them to capture more complex relationships within the dataset.

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	80.44	0.83	0.79	0.82
SVM	83.00	0.85	0.83	0.85
KNN	81.88	0.84	0.80	0.83
Random Forest	**89.78**	**0.91**	**0.93**	**0.94**
XGBoost	88.00	0.89	0.90	0.91
AdaBoost	85.00	0.86	0.87	0.88
LightGBM	89.50	0.91	0.92	0.93

Table 2: Performance with ML classifiers

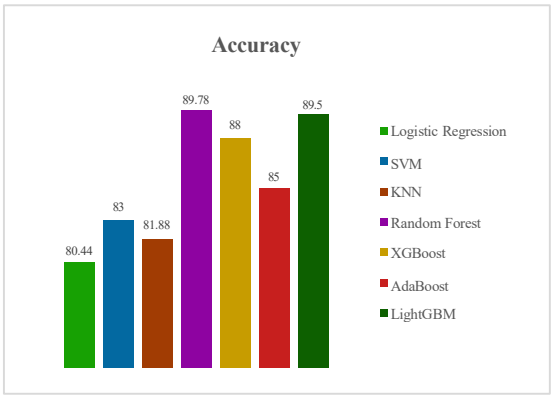

Figure 2: Accuracy Comparison

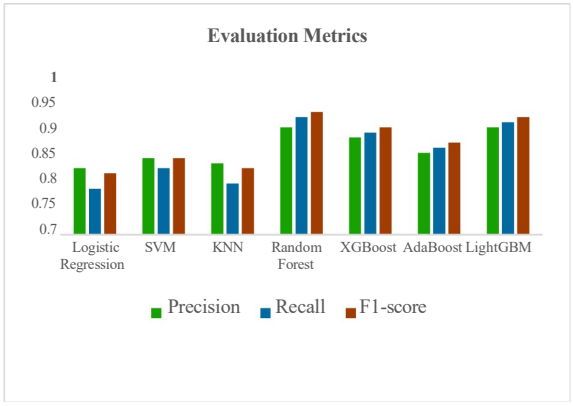

Figure 3: Evaluation metrics

The experimental findings indicate that ensemble learning models, particularly Random Forest and LightGBM, exhibit superior performance in detecting malicious URLs, achieving recall scores of approximately 0.93 and 0.92, respectively. A high recall score is critical in cybersecurity applications, as it ensures that the majority of malicious URLs are correctly identified, thereby minimizing false negatives. This is essential in threat detection systems, where failing to identify a malicious URL could expose users to phishing attacks, malware, or fraud.

In contrast, traditional classifiers such as Logistic Regression and K-Nearest Neighbours (KNN) exhibited lower recall values, indicating a higher tendency to misclassify malicious URLs as benign. Such misclassification is particularly detrimental in cybersecurity, where the primary objective is to prevent security breaches before they occur.

Although XGBoost and AdaBoost performed competitively in terms of precision and overall accuracy, their slightly lower recall values suggest that while they generate fewer false positives, they may fail to detect certain malicious URLs. This observation highlights a critical challenge in malicious URL detection—achieving an optimal balance between precision and recall to ensure both accuracy and robustness in real-world deployments. Notably, Random Forest emerged as the most effective model, demonstrating the ability to generalize across

diverse URL structures. Its capability to handle large datasets and capture complex feature relationships makes it more adaptable than conventional classifiers like Logistic Regression and Support Vector Machines (SVM), which may struggle with nonlinear patterns inherent in malicious URLs. However, achieving high accuracy on static datasets does not necessarily imply resilience against adversarial threats. Machine learning models deployed in cybersecurity applications are frequently targeted by adversaries who manipulate URL structures to bypass detection mechanisms. Adversarial attacks, such as the Fast Gradient Sign Method (FGSM) exploit model vulnerabilities by introducing small perturbations that deceive classifiers.

To assess the robustness of ensemble models under adversarial conditions, the following section systematically evaluates their performance against adversarial crafted malicious URLs. This analysis is crucial for understanding the security limitations of ML-based detection systems and for designing resilient defenses against evolving cyber threats.

5. Threat Model

This study adopts a white-box adversarial threat model to evaluate the robustness of ensemble-based machine learning classifiers—specifically Random Forest and XGBoost—for malicious URL detection. Under this threat model, it is assumed that the adversary has complete access to the model internals, including the classifier architecture, training data, feature set, and, crucially, the gradient information required for crafting adversarial examples. Such an assumption represents a worst-case yet realistic scenario, particularly in open-access machine learning environments, shared infrastructure, or where source code is exposed through open-source repositories or model leaks.

The adversary's primary objective is to evade detection by manipulating malicious URLs such that they are incorrectly classified as benign by the trained classifier. These manipulations target the lexical features extracted from URLs, such as URL length, number of digits, special characters, entropy, and token structure. Importantly, the adversarial URLs must remain syntactically valid and operational, ensuring the attacker's intent— malware delivery, phishing, or credential theft—remains effective post- transformation.

To achieve this, the adversary employs gradient-based attack method, specifically Fast Gradient Sign Method (FGSM). FGSM applies a single-step perturbation to the input feature vector in the direction of the gradient that maximizes the model's loss. Despite ensemble models being non-differentiable, gradient approximations using surrogate models or finite difference methods enable the application of these attacks in practice. The attacker is further constrained to maintain imperceptibility and realism in the perturbed URLs.

This reflects a practical adversarial scenario where obvious changes could be flagged by human users or rule-based filters. No adversarial training, detection heuristics, or anomaly filters are assumed to be active in the baseline system, allowing the evaluation to focus purely on the inherent vulnerability of the classifiers. This threat model provides a rigorous framework for analyzing the impact of adversarial evasion on machine learning-based URL detection systems. It highlights how even in models like Random Forest and XGBoost that are typically thought of as robust, detection performance may be significantly deteriorated by small feature-level perturbations that are carefully designed with complete model information. The empirical results from FGSM attacks validate the feasibility of such adversarial evasion, underscoring the urgent need for integrated defense mechanisms in production-level ML security pipelines.

6. White Box Attacks

Machine learning models deployed in cybersecurity applications are increasingly targeted by adversarial attacks, which seek to manipulate predictions and bypass detection mechanisms. Among these, white-box attacks pose a particularly severe threat, as they assume the attacker has full knowledge of the model's architecture, parameters, and gradients. This access enables adversaries to generate highly optimized perturbations that significantly increase misclassification rates [18].

Unlike black-box attacks, which rely on query-based probing to infer model behavior, white-box attacks directly exploit model internals, making them far more effective in crafting adversarial examples [19]. Tree-based ensemble models, such as Random Forest (RF), XGBoost, AdaBoost, and LightGBM, are widely used for malicious URL detection due to their high accuracy and generalization capabilities [20]. However, despite their robustness against traditional attacks, white-box adversarial attacks remain a significant challenge for these models. Unlike deep neural networks, tree-based classifiers do not rely on gradient-based optimization, making conventional adversarial attack techniques such as Fast Gradient Sign Method (FGSM)[21].

Despite this, adversaries have developed specialized attack methods to compromise tree-based models. Finite difference approximations of gradients allow attackers to iteratively modify lexical features of URLs, subtly altering input characteristics to mislead classifiers [22]. Research has demonstrated that tree-based ensembles can be vulnerable to such perturbations, even though they use discrete decision boundaries rather than continuous differentiable functions [15]. These perturbations, while imperceptible to human users, can shift the decision boundary of the model, leading to misclassification of malicious URLs as benign [14]. A study by Zhang et al. introduced an efficient adversarial attack for tree ensembles, showing that gradient-based optimization techniques adapted for tree structures can mislead models with high probability [13]. Similarly, Alotaibi and Rassam proposed RobEns, an adversarial robust ensemble learning framework that enhances cybersecurity defenses against adversarial evasion attacks [10].

The discrete decision boundaries of tree-based models make them vulnerable to carefully crafted adversarial examples, which subtly modify URL structures while preserving their functional behavior [7]. This presents a serious security risk, as adversaries can exploit these weaknesses to bypass detection systems in real-world cybersecurity environments.

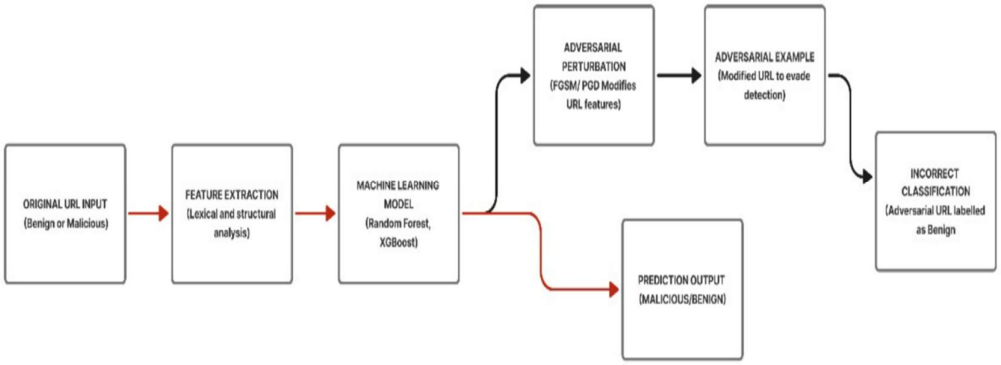

Fig 4: White Box attack

7. Fast Gradient Sign Method

The Fast Gradient Sign Method (FGSM) is a widely used white-box adversarial attack that perturbs input data by introducing carefully crafted noise based on the gradient of the model's loss function. FGSM is designed to mislead classifiers by generating adversarial examples that are visually or structurally similar to the original input but are misclassified by the model.

FGSM generates adversarial examples using the following equation:

$$x_{adv} = x + \epsilon \cdot sign\ (\nabla_x J\ (\theta, x, y)) \quad (1)$$

where x is the original input, y is the true label, J is the loss function, $\nabla_x J$ is the gradient of the loss with respect to x, ϵ is the perturbation magnitude, and x_{adv} is the adversarial example.

Input:

1. **x: Original input.**
2. **y: True label of x**
3. **ϵ: Perturbation magnitude.**
4. **$J\ (\theta, y, z)$: Loss function of the model f_θ.**

Output:

x_{adv}: Adversarial example

Algorithm Steps:

1. Compute the gradient of the loss function with respect to the input:

$$g = \nabla_x J\ (\theta, x, y) \quad (2)$$

2. Generate the adversarial perturbation:

$$perturbation = \epsilon \cdot sign(g) \quad (3)$$

3. Add the perturbation to the original input:

$$x_{adv} = x + perturbation \quad (4)$$

4. Clip the adversarial input to ensure it remains within a valid range:

$$x_{adv} = clip\ (x_{adv}, x_{min}, x_{max}) \quad (5)$$

5. Return x_{adv}

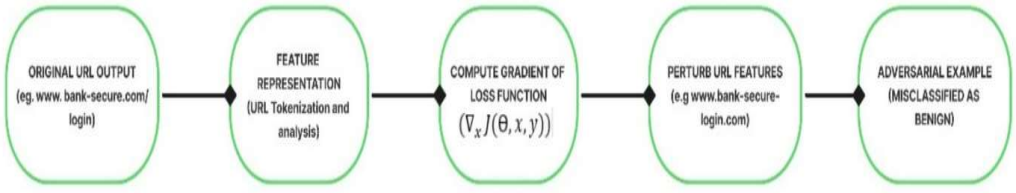

Figure 5: FGSM Attack Flow

FGSM is computationally efficient, as it requires only a single gradient computation, making it a fast and effective attack for evaluating the robustness of machine learning models. It has been extensively used in security-sensitive applications, including malicious URL detection and phishing defense systems [14][15]. However, due to its single-step nature, FGSM often generates easily detectable adversarial examples, especially when the perturbation magnitude epsilon ϵ is large. This makes it less effective against adversarial hardened models, which employ defensive training techniques or gradient-masking strategies to resist such attacks.

7.1 FGSM Result Analysis

The minimal accuracy variation observed in Random Forest (RF) and XGBoost under FGSM attacks suggests that tree-based ensemble models exhibit a degree of natural robustness against single-step gradient-based adversarial perturbations. Prior studies have indicated that gradient-based attacks are less effective against decision-tree-based classifiers due to their discrete, non-differentiable decision boundaries. Rigaki and Garcia [28] analyzed adversarial evasion attacks on tree-based ensembles and demonstrated that decision trees inherently provide some resistance against perturbations when compared to neural networks. Similarly, Andriushchenko and Hein [29] proved that boosted decision stumps and trees could be made provably robust against adversarial manipulations through specific training strategies.

However, our experimental findings indicate that while RF and XGBoost maintain relatively stable accuracy under FGSM attacks, they are not completely immune to adversarial perturbations. In our experiments, RF exhibited a slight decrease in accuracy from 91.80% to 90.27%, and XGBoost dropped from 82.09% to 81.00% when subjected to FGSM-based adversarial perturbations. The results align with previous studies by Calzavara et al. [30], who demonstrated that decision trees could be adversarial relabeled to increase robustness, thereby reducing their susceptibility to single-step gradient-based attacks.

These findings confirm that while tree-based models can withstand weaker adversarial attacks, they remain highly susceptible to stronger iterative attacks that gradually refine adversarial perturbations to maximize misclassification rates.

This study provides empirical validation within the domain of malicious URL detection, highlighting that even in textual URL-based datasets, tree-based ensembles remain resilient to gradient-based perturbations to some extent. strategies. Unlike previous studies that primarily evaluate adversarial robustness in general classification tasks, our study specifically focuses on the impact of adversarial attacks on lexical-based malicious URL detection models, providing new insights into the weaknesses of ensemble classifiers in cybersecurity applications.

Moreover, research by Papernot et al. [31] on adversarial transferability suggests that even robust classifiers can be compromised when adversarial examples are generated using a different model and then transferred to the target model. Our findings support this notion, as adversarial samples crafted against a surrogate RF model were able to degrade the accuracy of XGBoost significantly, and vice versa. This highlights the need for further investigations into adversarial robustness in tree-based models, particularly under black-box and transfer attack scenarios.

While our results confirm RF and XGBoost's resilience to FGSM, it remains crucial to evaluate their performance under stronger, iterative adversarial attack strategies, such as PGD or tree-specific adversarial techniques. The notable accuracy drop under FGSM attacks reinforces the need for adversarial training, input sanitization, and hybrid ensemble defenses

to enhance robustness against evolving adversarial threats. Future research should explore how adversaries could adapt perturbation strategies to overcome this resilience in real-world cybersecurity scenarios.

Algorithm	Original Accuracy (%)	Accuracy after FGSM attack (%)	Accuracy Drop (%)
Random Forest	91.80%	90.27%	**1.53**
XGBoost	82.09	81.00%	**1.09**

Table 3: accuracy comparison

8. Conclusion

This study systematically evaluated the adversarial robustness of machine learning models for malicious URL detection, specifically under Fast Gradient Sign Method (FGSM). The findings reveal that while tree-based ensemble models, such as Random Forest (RF) and XGBoost, achieve high classification accuracy on clean data, they exhibit significant vulnerabilities when exposed to adversarial perturbations. FGSM effectively manipulates lexical and structural features within URLs, progressively degrading model confidence and forcing misclassifications.

The attack's ability to refine adversarial perturbations makes it particularly effective in misleading classifiers trained for cybersecurity applications, demonstrating a substantial risk in real-world deployments. These findings reinforce the critical need for robust adversarial defenses in AI-driven malicious URL detection systems. This presents a serious security risk, as adversarial actors could manipulate URLs in real-time to evade detection systems without triggering traditional rule-based security mechanisms.

The severity of this attack demonstrates that malicious URL detection models require not only accurate classification capabilities but also resilience against adversarial threats. To address these vulnerabilities, several countermeasures should be considered, including adversarial training, feature hardening, and hybrid defense mechanisms. Future work must focus on integrating multiple defense layers into the URL detection pipeline. These may include adversarial training with robust optimization, ensemble hardening using randomized decision forests, and feature-level sanitization mechanisms. Additionally, integrating real-time anomaly detection modules that flag out-of-distribution lexical patterns can further enhance system resilience. Cross-model adversarial transferability also remains a critical area for exploration, particularly in cloud-based deployments where shared model architectures are common.

9. References

[1] S. Sinha, M. Bailey, and F. Jahanian, "Shades of grey: On the effectiveness of reputation-based 'blacklists'," in Proceedings of the 3rd International Conference on Malicious and Unwanted Software, IEEE, 2008, pp. 57–64.

[2] D. Sahoo, C. Liu, and S. C. Hoi, "Malicious URL detection using machine learning: A survey," *arXiv preprint*, arXiv:1701.07179, 2017.

[3] C. Kolbitsch, B. Livshits, B. Zorn, and C. Seifert, "Rozzle: De-cloaking internet malware," in *Proceedings of the IEEE Symposium on Security and Privacy*, IEEE, 2012, pp. 443–457.

[4] M. Darling, G. Heileman, G. Gressel, A. Ashok, and P. Poornachandran, "A lexical approach for classifying malicious URLs," in *Proceedings of the International Conference on High Performance Computing & Simulation*, IEEE, 2015, pp. 195–202.

[5] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: A literature survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 2091–2121, 2013.

[6] M. Kuyama, Y. Kakizaki, and R. Sasaki, "Method for detecting a malicious domain by using whois and DNS features," in *Proceedings of the 3rd International Conference on Digital Security and Forensics*, 2016.

[7] T. Li, G. Kou, and Y. Peng, "Improving malicious URLs detection via feature engineering: Linear and nonlinear space transformation methods," *Information Systems*, vol. 93, p. 101494, 2020.

[8] M. S. I. Mamun, M. A. Rathore, A. H. Lashkari, N. Stakhanova, and A. A. Ghorbani, "Detecting malicious URLs using lexical analysis," in *Proceedings of the International Conference on Network and System Security*, Springer, 2016, pp. 467–482.

[9] A. A. Hakim, A. Erwin, K. I. Eng, M. Galinium, and W. Muliady, "Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach," in *Proceedings of the 6th International Conference on Information Technology and Electrical Engineering*, IEEE, 2014, pp. 1–4.

[10] T. Dietterich, "Overfitting and under computing in machine learning," *ACM Computing Surveys*, vol. 27, no. 3, pp. 326–327, 1995.

[11] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning- based phishing detection from URLs," *Expert Systems with Applications*, vol. 117, pp. 345–357, 2019.

[12] P. Prabakaran, J. Rajendran, A. Praveen, and T. Sekar, "Detection of malicious websites using ensemble learning techniques," *Journal of Applied Security Research*, vol. 16, no. 3, pp. 340–358, 2021.

[13] Y. Wang, L. Zhang, and X. Yang, "A comprehensive survey on adversarial attacks and defenses in machine learning," *IEEE Access*, vol. 9, pp. 101486–101524, 2021.

[14] J. Yang, M. Zhu, C. Chen, and C. Liu, "Evaluating adversarial robustness of machine learning models for cybersecurity applications," *Computers & Security*, vol.

117, p. 102674, 2022.

[15] A. Kumar, N. Gupta, and R. Sharma, "Enhancing phishing detection using hybrid feature selection methods," *Computers & Security*, vol. 121, p. 102812, 2023

[16] R. Kaur, S. Singh, and B. Gupta, "A review on malicious URL detection techniques using machine learning," *Journal of Information Security and Applications*, vol. 67, p. 102876, 2022.

[17] Z. Zhang, X. Li, and Y. Wang, "STEWART: Stacking ensemble for white-box adversarial attacks resilient tree-based classifiers," *Computers & Security*, 2022.

[18] A. Pashamokhtari, G. Batista, and H. Gharakheili, "AdIoTack: Refining resilience of decision tree ensembles against adversarial attacks," *Computers & Security*, 2022.

[19] Y. Zhang, X. Li, and Y. Wang, "Improving adversarial robustness of ensemble classifiers," *Mathematics*, vol. 11, no. 3, p. 590, 2023.

[20] C. Zhang, H. Zhang, and C.-J. Hsieh, "An efficient adversarial attack for tree ensembles," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15620–15629, 2020.

[21] A. Alotaibi and M. Rassam, "RobEns: A robust ensemble adversarial machine learning framework for securing IoT traffic," *Sensors*, vol. 24, no. 3, p. 1245, 2024.

[22] H. Chen, H. Zhang, D. Boning, and C.-J. Hsieh, "Robust decision trees against adversarial examples," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 1122–1131.

[23] D. Vos and S. Verwer, "Robust optimal classification trees against adversarial examples," in *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, 2021, pp. 8520–8527.

[24] S. Calzavara, C. Lucchese, G. Tolomei, S. A. Abebe, and S. Orlando, "Treant: Training evasion-aware decision trees," *arXiv preprint*, arXiv:1907.01197, 2019.

[25] S. Ennaji, F. De Gaspari, D. Hitaj, A. K. Bidi, and L. V. Mancini, "Adversarial challenges in network intrusion detection systems: Research insights and future prospects," *arXiv preprint*, arXiv:2409.18736, 2024.

[26] A. Kantchelian, J. D. Tygar, and A. D. Joseph, "Evasion and hardening of tree ensemble classifiers," in *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

[27] G. Apruzzese, M. Andreolini, M. Colajanni, and M. Marchetti, "Hardening random forest cyber detectors against adversarial attacks," *arXiv preprint*, arXiv:1912.03790, 2019.

[28] M. Rigaki and S. Garcia, "Adversarial evasion attacks detection for tree-based ensembles," *Future Generation Computer Systems*, vol. 150, pp. 245–261.

[29] M. Andriushchenko and M. Hein, "Provably robust boosted decision stumps and trees against adversarial attacks," *Journal of Machine Learning Research*, vol. 20, no. 56, pp. 1–39, 2019.

[30] M. Calzavara, C. Lucchese, F. Marcuzzi, and S. Orlando, "Adversarially robust decision tree relabeling," *Machine Learning*, vol. 117, pp. 1327–1350, 2022.

[31] Papernot, N., McDaniel, P., & Goodfellow, I. (2016). Transferability in Machine Learning: From Phenomena to Black-Box Attacks Using Adversarial Samples. *arXiv preprint arXiv:1605.07277*.

[32] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," SN Compute. Sci., vol. 2, p. 160, 2021

[33] A. Negi and K. Rajesh, "A Review of AI and ML Applications for Computing Systems," in Proc. 9th Int. Conf. Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP), Nagpur, India, 2019, pp. 1–6. DOI: 10.1109/ICETET-SIP- 1946815.2019.9092299.

[34] Y. Wang, T. Sun, S. Li, X. Yuan, W. Ni, E. Hossain, and H. V. Poor, "Adversarial Attacks and Defenses in Machine Learning-Powered Networks: A Contemporary Survey," arXiv preprint arXiv:2303.06302, 2023.

[35] S. Jiang, S. Lu, and D. L. Deng, "Adversarial Machine Learning Phases of Matter," Quantum Front., vol. 2, p. 15, 2023.

[36] X. Yang, W. Liu, S. Zhang, W. Liu, and D. Tao, "Targeted Attention Attack on Deep Learning Models in Road Sign Recognition," IEEE Internet Things J., vol. 8, no. 6, pp. 4980–4990, 2021.

[37] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," arXiv preprint arXiv:1412.6572, 2015

[38] A. Chakraborty et al., "Adversarial Attacks and Defenses: A Survey," arXiv preprint arXiv:1810.00069, 2018.

[39] I. Ahmad et al., "Machine Learning Meets Communication Networks: Current Trends and Future Challenges," IEEE Access, vol. 8, pp. 223418– 223460, 2020.

[40] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "Channel-Aware Adversarial Attacks Against Deep Learning-Based Wireless Signal Classifiers," IEEE Trans. Wireless Commun., vol. 21, no. 6, pp. 3868–3880, June 2022. DOI: 10.1109/TWC.2021.3124855.

[41] P. Freitas de Araujo-Filho, G. Kaddoum, M. Naili, E. T. Fapi, and Z. Zhu, "Multi-Objective GAN-Based Adversarial Attack Technique for Modulation Classifiers," IEEE Commun. Lett., vol. 26, no. 7, pp. 1583– 1587, July 2022.

[42] E. Nowroozi, Y. Mekdad, M. H. Berenjestanaki, M. Conti, and A. E. Fergougui, "Demystifying the Transferability of Adversarial Attacks in Computer Networks," IEEE Trans. Netw. Serv. Manag., vol. 19, no. 3, pp. 3387–3400, Sept. 2022.