Surrogate-Guided Adversarial Attacks: Enabling White-Box Methods in Black-Box Scenarios

Dimitrios Christos Asimopoulos
MetaMind Innovations, Kozani, Greece
Department of Information and Electronic Engineering, *International Hellenic University*, Thessaloniki, Greece
Email: dasimopoulos@metamind.gr, dimiasim3@ihu.gr

Panagiotis Radoglou-Grammatikis
K3Y Ltd, Sofia, Bulgaria
Department of Electrical and Computer Engineering, *University of Western Macedonia*, Kozani, Greece
Email: pradoglou@k3y.bg, pradoglou@uowm.gr

Panagiotis Fouliras
Department of Applied Informatics, *University of Macedonia*, Thessaloniki, Greece
Email: pfoul@uom.edu.gr

Konstandinos Panitsidis
Department of Management Science & Technology, *University of Western Macedonia*, Kozani, Greece
Email: kpanytsidis@uowm.gr

Georgios Efstathopoulos
MetaMind Innovations, Kozani, Greece
Email: gefstathopoulos@metamind.gr

Thomas Lagkas
Department of Computer Science, *Democritus University of Thrace*, Kavala, Greece
Email: tlagkas@cs.duth.gr

Vasileios Argyriou
Department of Networks and Digital Media, *Kingston University London*, Penrhyn Road, UK
Email: vasileios.argyriou@kingston.ac.uk

Igor Kotsiuba
Durham University Business School, Millhill Ln, UK
Email: igor.kotsiuba@durham.ac.uk

Panagiotis Sarigiannidis
Department of Electrical and Computer Engineering, *University of Western Macedonia*, Kozani, Greece
Email: psarigiannidis@uowm.gr

***Abstract*—Adversarial attacks pose significant threats to machine learning models, with white-box attacks such as Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Basic Iterative Method (BIM) achieving high success rates when model gradients are accessible. However, in real-world scenarios, direct access to model internals is often restricted, necessitating black-box attack strategies that typically suffer from lower effectiveness. In this work, we propose a novel approach to transform white-box attacks into black-box attacks by leveraging state-of-the-art surrogate models, including Multi-Layer Perceptrons (MLP) and XGBoost (XGB). Our method involves training a surrogate model to mimic the decision boundaries of an inaccessible target model using pseudo-labeling, thereby enabling the application of gradient-based white-box attacks in a black-box setting. We systematically compare our approach against conventional black-box attacks, such as Zero Order Optimization (ZOO), evaluating their effectiveness in terms of attack success rates, transferability, and computational efficiency. The results demonstrate that surrogate-assisted attacks perform as good as standard black-box methods, bridging the performance gap between white-box and black-box adversarial attacks. This study highlights the power of surrogate models in enhancing adversarial transferability and provides insights into the robustness of different machine learning architectures against adversarial threats.**

***Index Terms*—Adversarial attacks, white-box, Black-box, evasion, transferability, surrogate-model**

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450 (AI4CYBER). Disclaimer: Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.

I. Introduction

Adversarial attacks pose significant threats to machine learning (ML) models, particularly in applications such as computer vision, natural language processing, and cybersecurity. These attacks exploit model vulnerabilities by introducing small perturbations to input data, leading to high-confidence misclassifications. White-box attacks, including Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Basic Iterative Method (BIM), achieve high success rates by leveraging full access to the model's architecture and gradients. However, in real-world scenarios, such access is typically unavailable, making black-box attack strategies necessary. These approaches rely on query-based methods, such as Zero Order Optimization (ZOO), or transferability of

adversarial examples from surrogate models, though often with reduced effectiveness.

Black-box attacks face key limitations due to limited knowledge of the target model's decision boundary, high query demands, and unreliable transferability—particularly when targeting ensemble models like XGBoost. To address these challenges, this work proposes a surrogate-based black-box attack framework designed to improve adversarial transferability without direct gradient access. A neural network surrogate is trained using pseudo-labels from the target XGBoost model, allowing the application of gradient-based white-box attacks in a black-box context.

The main contributions of this paper are summarized as follows:

- **Surrogate-Based Black-Box Framework:** A structured attack methodology using a neural network surrogate trained via pseudo-labeling to enable effective adversarial generation against XGBoost.
- **White-Box Attack Adaptation:** Application of white-box attacks in black-box scenarios through surrogate-assisted transfer.
- **Comparative Evaluation:** Systematic comparison between the proposed surrogate-based approach and the ZOO black-box attack, focusing on success rates, transferability, and computational efficiency.

This approach aims to bridge the gap between white-box and black-box adversarial attacks, enhancing attack effectiveness against non-differentiable models like XGBoost. By leveraging surrogate models, it becomes possible to approximate the target model's behavior and generate transferable adversarial examples with improved success rates. Furthermore, the framework reduces the dependency on excessive query counts, making the attack process more scalable and practical for real-world adversarial testing.

The rest of the paper is organized as follows. Section II presents a background and similar works in this field. Section III, provides the architecture of the proposed work alongside the methodology. Next, section IV provides the dataset and the metrics used, section V focuses on the evaluation analysis and experimental results, while VI concludes this paper.

II. Related Work

Deep neural networks (DNNs) have achieved state-of-the-art performance across various domains but remain inherently vulnerable to adversarial examples—carefully crafted perturbations that cause incorrect predictions. This vulnerability poses a significant challenge, particularly in black-box settings where attackers have no access to the model's internal parameters and must rely solely on observed input-output behavior. To overcome this, many attack strategies leverage surrogate models to approximate the decision boundaries of the target model and generate transferable adversarial examples that maintain effectiveness across model boundaries [1].

Adversarial attacks are typically classified into white-box and black-box categories. White-box attacks, such as FGSM, PGD, and Carlini & Wagner (C&W), assume full access to model architecture and gradients, enabling precise, gradient-based perturbations. In contrast, black-box attacks rely on the transferability phenomenon, whereby adversarial examples generated from a surrogate model can deceive a different, unseen model. This transferability underpins the majority of black-box attacks and has motivated extensive research on improving the fidelity of surrogate models to approximate target behaviors more effectively [2].

To address the surrogate-to-target mismatch, several methods have emerged. One notable advancement is the Lipschitz Regularized Surrogate (LRS), which introduces Lipschitz continuity constraints to smooth the surrogate model's loss surface. This regularization enhances the generalizability of perturbations, significantly increasing black-box attack success rates against both standard and adversarially-trained models [3].

Beyond algorithmic contributions, frameworks for systematic robustness evaluation have also gained traction. The Adversarial Attack Generator (AAG) proposed in [4] is a modular platform designed to evaluate ML/DL models' resilience to attacks in critical infrastructure domains. AAG uses the CICFlowMeter parser to extract features from OCPP-based traffic and applies a variety of adversarial techniques—including FGSM, JSMA, PGD, and C&W—via its attack engine. Its evaluation module benchmarks different classifiers, revealing performance degradation under adversarial conditions and emphasizing the need for adaptive defense strategies.

Further emphasizing this need, [5] investigated the robustness of AI-enabled Intrusion Detection Systems (IDS) in the energy sector, specifically targeting the IEC 60870-5-104 protocol. They employed both gradient-based attacks (FGSM) and data-driven synthetic perturbations generated via Conditional Tabular GANs (CTGANs). Their findings revealed a substantial drop in detection accuracy across several classifiers, including Random Forest, XGBoost, and MLP, when subjected to adversarial inputs—highlighting the limitations of conventional IDSs under adversarial threat.

One key advancement in enhancing white-box attack transferability to black-box settings is the work by Inkawhich et al. [6], which introduces a feature-space attack strategy that perturbs internal activation patterns instead of just final predictions. By manipulating features across different hierarchy levels in the network, the attack exploits deeper representations shared across various model architectures, leading to improved transferability.

In a similar vein, Wu et al. proposed the Skip Gradient Method (SGM) [7], which strategically suppresses gradient flow through residual connections during the attack generation process. This method mitigates the overfitting of perturbations to the specific architectural shortcuts of the surrogate model, improving generalization to unseen architectures. Notably, when used in combination with iterative attacks like I-FGSM and MI-FGSM, SGM consistently enhances black-box success rates, particularly in models with deep residual blocks such as ResNets and DenseNets.

Wang et al. [8] contributed another dimension to white-

box to black-box adaptation through their Variance Tuning Method (VTM). Instead of modifying the surrogate model itself, VTM focuses on controlling the variance of gradient updates during adversarial example generation. By balancing the exploration and exploitation trade-off in the perturbation process, the method avoids overfitting to the surrogate model's landscape. The result is a more diverse set of perturbations that maintain their adversarial properties across target models.

Dong et al. introduced the Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [9], which integrates momentum into the iterative white-box attack process. This technique helps the perturbation trajectory escape local optima associated with the surrogate's loss surface, leading to more stable and generalized adversarial examples. MI-FGSM has been shown to significantly outperform standard iterative methods in black-box scenarios, especially when targeting commercial-grade classifiers and adversarially trained networks.

Building on this momentum approach, Lin et al. [10] proposed the Nesterov Iterative FGSM (NI-FGSM), which applies Nesterov Accelerated Gradient (NAG) to anticipate the future direction of gradient descent during adversarial optimization. This forecasting behavior improves the attack's convergence and helps craft perturbations that are both more transferable and less perceptible. When benchmarked on large-scale datasets like ImageNet, NI-FGSM achieved higher black-box success rates and exhibited better performance under ensemble-targeting scenarios.

Taken together, these works demonstrate the evolving sophistication of adversarial attacks and the critical role of surrogate modeling in black-box threat scenarios. However, despite these advancements, challenges remain in improving cross-model transferability, adversarial generalization, and robustness evaluation under real-world constraints. These gaps motivate the need for improved methods that unify model realism, attack efficacy, and systematic evaluation.

III. Methodology

A. Methodology Analysis

The proposed methodology addresses one of the key limitations of adversarial machine learning in black-box settings: the inability to apply gradient-based attacks directly on non-differentiable models such as XGBoost. Traditional black-box attacks often suffer from high query costs or unreliable transferability. To overcome these challenges, this work introduces a surrogate-assisted attack framework that enables the use of efficient white-box attacks in black-box scenarios by approximating the decision boundary of the target model through pseudo-labeling.

The approach offers two main advantages. First, it reduces the dependency on excessive query-based optimization by relying on a substitute model trained to mimic the behavior of the target classifier. Second, it enhances the transferability of adversarial examples by aligning perturbation directions with the surrogate model's learned gradients. By systematically comparing direct query-based methods like Zero Order Optimization (ZOO) with surrogate-assisted attacks, the methodology provides insights into the effectiveness of white-box attacks when adapted to black-box environments. This analysis helps quantify not only attack success rates but also the efficiency and scalability of the proposed strategy.

B. Problem Definition

In black-box adversarial attack scenarios, attackers lack access to the target model's internal parameters and gradients and can only query the model to observe its outputs. Formally, given a dataset $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$, where X_i denotes the input features and y_i the corresponding labels, the target black-box classifier f_{bb} maps inputs to predictions according to $f_{\text{bb}} : X \to y$.

In such settings, direct gradient-based attacks like FGSM [11], PGD [12], or C&W [13] cannot be applied because models like XGBoost, Random Forest, and Decision Trees do not expose differentiable structures. Therefore, an alternative approach is required to enable effective adversarial perturbation generation.

C. Surrogate Model Training

To bypass the non-differentiability of the target model, we introduce a surrogate learning phase where a differentiable substitute model f_{sub}, parameterized by θ, is trained to approximate the decision boundary of f_{bb}. The surrogate model employed in this work is a Multi-Layer Perceptron (MLP), designed as a fully connected neural network with four layers. The architecture integrates ReLU activations, batch normalization, L2 regularization, and dropout mechanisms to ensure stability and prevent overfitting [14].

The MLP architecture begins with an input layer of 256 neurons, followed by two hidden layers of 128 and 64 neurons, progressively reducing the feature space. The output layer applies a softmax activation function to support multi-class classification using pseudo-labels provided by the target model. The training process uses the Adam optimizer [15] with a learning rate of 0.005, optimizing the sparse categorical cross-entropy loss over 50 epochs with a batch size of 64. Validation is performed on a hold-out dataset to monitor generalization, while dropout at a 50% rate mitigates overfitting during training.

Rather than using true labels, the surrogate is trained on pseudo-labels generated by querying the target black-box model, minimizing the following loss function:

$$\min_{\theta} \sum_{i=1}^{N} \mathcal{L}\left(f_{\text{sub}}(X_i;\theta), f_{\text{bb}}(X_i)\right), \tag{1}$$

where $\mathcal{L}$ denotes the classification loss, typically cross-entropy. This pseudo-labeling strategy allows the surrogate to approximate the decision regions of the target model, providing the necessary gradients for subsequent adversarial attacks.

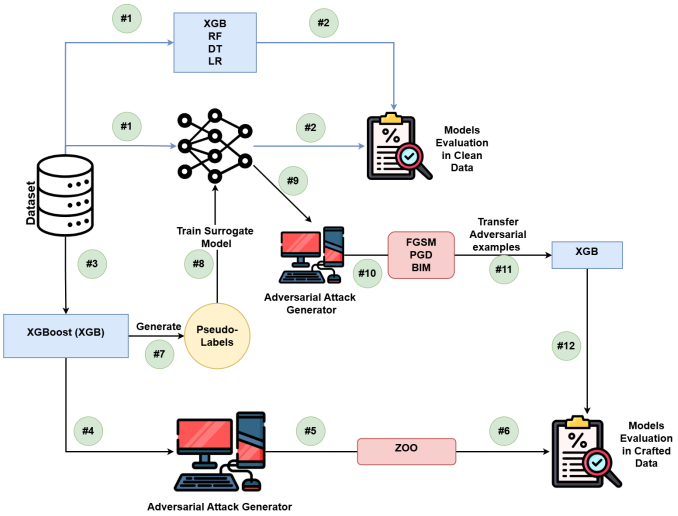

Fig. 1. Workflow for surrogate-assisted black-box adversarial attacks.

D. White-Box Attack Adaptation

After training the surrogate model, standard white-box adversarial attacks, including FGSM, PGD, and BIM, can be applied. These methods compute perturbations using gradients from the surrogate model's loss function. An adversarial example X_{adv} is generated by perturbing a clean input X in the direction of the loss gradient:

$$X_{\text{adv}} = X + \epsilon \cdot \text{sign}\left(\nabla_X \mathcal{L}(f_{\text{sub}}(X), y)\right), \tag{2}$$

where ϵ controls the perturbation magnitude and $\nabla_X \mathcal{L}$ denotes the gradient of the loss function with respect to the input. Although crafted on the surrogate, these adversarial examples aim to remain effective when transferred to the original black-box target model.

E. Black-Box Evaluation and Attack Workflow

The overall methodology is illustrated in Fig. 1. The process begins by establishing baseline performance using clean input data. Direct query-based attacks, such as ZOO, are applied as a baseline for comparison. Following this, the surrogate model is trained on pseudo-labeled data generated from the target classifier. White-box attacks are then applied to the surrogate, and the crafted adversarial examples are transferred back to the target model for evaluation.

This approach enables the assessment of both the direct black-box attack and the surrogate-assisted strategy, offering comparative insights into their respective attack success rates, transferability performance, and computational costs. By bridging white-box methods with black-box scenarios, the methodology contributes to a more effective and scalable framework for adversarial testing against non-differentiable models.

The results from both black-box and white-box attack strategies are compared to assess their effectiveness, and the findings are ultimately communicated to the system user for further analysis and decision-making.

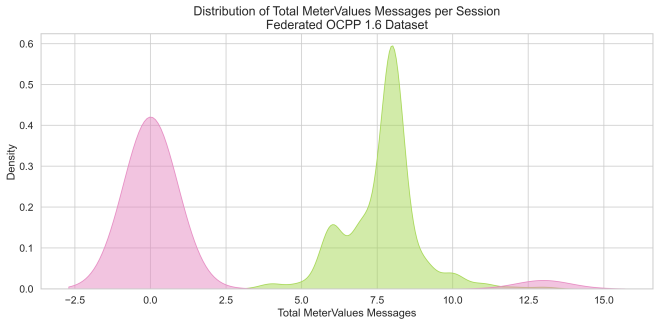

Fig. 2. Kernel Density Estimation (KDE) plot of the flow_total_ocpp16_metervalues feature from the Federated OCPP 1.6 Intrusion Detection Dataset

IV. Experiment Setup

The experimental results were conducted on a MacBook Air M2 equipped with an Apple M2 chip, 8GB of unified memory, and 256GB of SSD storage. Despite its compact and energy-efficient architecture, the M2 chip provides significant computational power through its integrated GPU and Neural Engine, enabling efficient execution of machine learning and deep learning tasks. The preferred deep learning framework for this experiment was TensorFlow, selected for its optimization on Apple Silicon and seamless integration with macOS, ensuring smooth execution of adversarial attack evaluations.

A. Dataset

The dataset utilized in this study is part of the OCPP (Open Charge Point Protocol) Dataset, which was parsed using CICFlowMeter to extract network flow statistics.

The dataset utilized in this study is the Federated OCPP 1.6 Intrusion Detection Dataset [16], which contains network traffic and labeled cyberattack data targeting the Open Charge Point Protocol (OCPP) 1.6. This dataset is specifically designed to support AI-driven Intrusion Detection Systems (IDS) and includes various adversarial scenarios relevant to electric vehicle (EV) charging infrastructure. The recorded attacks include Charging Profile Manipulation, Denial of Charge, Heartbeat Flooding DoS, and Unauthorized Access, among others. This dataset comprises various network flow features recorded in PCAP CSV format, providing detailed insights into network behavior.

To illustrate the internal behavioral structure of the dataset, Fig.2 shows the distribution of MeterValues messages per session across all recorded labels. This feature reflects how frequently energy consumption metrics are reported in EV charging sessions and is particularly sensitive to disruptions caused by cyberattacks.The dataset includes both normal/benign traffic and multiple types of cyberattacks such as FDI Charging Profile, DOC ID Tag, DOS Flooding Heartbeat, and DOS Flooding EVCS Rejected attacks. Preprocessing steps involved feature engineering to drop nonpredictive features, identifying and handling null values, label encoding of target values, and standard scaling of the features to ensure they are on

a common scale. For this study, the dataset is leveraged to construct an adversarial dataset by applying white-box adversarial attack methods on surrogate models to simulate black-box attack scenarios and compare them with black-box attacks. The generated adversarial samples are used to assess the robustness of IDS models against adaptive adversarial threats. The combination of raw network traffic (PCAP) and structured flow statistics (CSV) provides a rich foundation for evaluating the efficacy of adversarial defense mechanisms in real-world EV charging environments.

B. Evaluation Metrics

1) Accuracy: The metric of accuracy measures the proportion of correct classifications in relation to the total instances. This evaluation metric is considered appropriate when the training dataset is balanced, meaning it contains an equal number of instances for all classes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

where:

$$TP \rightarrow \text{True Positives}$$
$$TN \rightarrow \text{True Negatives}$$
$$FP \rightarrow \text{False Positives}$$
$$FN \rightarrow \text{False Negatives}$$

2) True Positive Rate: TPR represents the fraction of actual intrusion instances that were correctly identified as intrusions.

$$TPR = \frac{TP}{TP + FN} \tag{4}$$

3) False Positive Rate: FPR indicates the proportion of normal instances that were incorrectly classified as cyberattacks, reflecting the balance between the accurate identification of normal instances and the occurrence of false alarms.

$$FPR = \frac{FP}{FP + FN} \tag{5}$$

4) F1 Score: The F1 score is a metric that captures the balance between true positive rate (TPR) and precision. Precision is defined as the ratio of true positives to the sum of true positives and false positives.

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{6}$$

5) Accuracy Drop: Calculates the difference between the accuracy before the attack and the accurace after the attack.

$$\Delta A = A_{\text{before}} - A_{\text{after}} \tag{7}$$

where:

- A_{before} is accuracy before the attack.
- A_{after} is accuracy after the attack.

6) Transferability Score: The transferability score T quantifies how effectively adversarial examples generated using the substitute model transfer to the black-box model. It is defined as:

$$T = \frac{\sum_{i=1}^{N} \mathbb{1}[f_{\text{bb}}(X_{\text{adv},i}) \neq y_i]}{\sum_{i=1}^{N} \mathbb{1}[f_{\text{sub}}(X_{\text{adv},i}) \neq y_i]} \tag{8}$$

where N is the total number of adversarial examples, $X_{\text{adv},i}$ is the i-th adversarial sample, and y_i is its true label. The functions $f_{\text{bb}}(\cdot)$ and $f_{\text{sub}}(\cdot)$ denote the black-box and substitute models, respectively. The indicator function $\mathbb{1}[\cdot]$ returns 1 if the condition is true and 0 otherwise. The score T reflects the ratio of successful attacks on the black-box model relative to those on the substitute model, indicating the transferability of the adversarial examples.

V. Experimental Results

A. Evaluation of Models in clean data

The experimental results are conducted based on the workflow of the proposed system. Table I presents the performance evaluation of the XGBoost model on the clean dataset before the application of adversarial attacks. The models exhibit high accuracy and F1-scores, indicating their effectiveness in classifying the dataset under normal conditions. The XGBoost model achieves an accuracy of 0.9335 on the clean dataset, with a corresponding F1-score of 0.9317. The True Positive Rate (TPR) matches the accuracy at 0.9335, indicating consistent performance in correctly classifying positive instances. The False Positive Rate (FPR) remains low at 0.0166, demonstrating good specificity. These results establish the baseline performance of the XGBoost model prior to the application of adversarial attacks, serving as a reference point for evaluating the model's robustness and vulnerability under adversarial perturbations.

TABLE I
Evaluation Metrics of XGBoost Model on Clean Dataset (Before Adversarial Attacks)

Metric	Accuracy	F1-score	TPR (Recall)	FPR
Score	0.9335	0.9317	0.9335	0.0166

B. Evaluation of the Black-Box Model Under Adversarial Attacks

The next step of our methodology focuses on evaluating the XGBoost model after applying the ZOO black-box attack. Table II summarizes the evaluation results for XGBoost under this attack scenario, providing key performance metrics as described in Section IV-B.

The results clearly demonstrate a significant degradation in model performance under adversarial perturbations. The accuracy of XGBoost drops sharply from its baseline clean performance of 0.9335 to 0.5259 after the ZOO attack. The F1-score and TPR similarly decline to 0.5134 and 0.5259, respectively, further confirming the effectiveness of the attack

in disrupting the model's predictive capabilities. Meanwhile, the FPR increases to 0.1185, reflecting a notable rise in misclassification of negative instances.

TABLE II
EVALUATION RESULTS OF THE XGBOOST MODEL AFTER APPLYING THE ZOO BLACK-BOX ATTACK.

Metric	ZOO (XGBoost)
Accuracy	0.5259
F1-score	0.5134
TPR	0.5259
FPR	0.1185
Accuracy Drop	0.4076

These findings highlight the susceptibility of XGBoost to black-box adversarial attacks, even when only query access is available. The considerable drop in accuracy underscores the vulnerability of such models in real-world deployment scenarios where direct gradient information is inaccessible.

C. Evaluation of White-Box Attacks on the Surrogate Model

After evaluating the impact of black-box attacks, we assessed the robustness of the surrogate model under white-box adversarial attacks using epsilon equal to 0.7 in order to have a big impact and have noteable results. Specifically, FGSM, PGD, and BIM were applied to surrogate models trained using. The results are detailed in Table III. The XGBoost surrogate model exhibited moderate resilience against adversarial perturbations, with its accuracy dropping from its clean state to 59.69% after FGSM and slightly lower to 59.38% under BIM. While the accuracy drop ranged between 33.67% and 33.98%, the model retained some robustness against adversarial attacks. However, transferability was significantly high for FGSM (99.81%), indicating that simple gradient-based attacks were highly effective in perturbing the model. This effectiveness decreased when iterative methods like PGD (89.12%) and BIM (69.48%) were applied, showing that XGBoost maintained a degree of robustness against more sophisticated white-box attacks.

TABLE III
EVALUATION RESULTS AFTER APPLYING WHITE-BOX METHODS TO THE SURROGATE MODEL USING XGBOOST.

	Epsilon = 0.7		
	FGSM	**PGD**	**BIM**
Accuracy	0.5969	0.5953	0.5938
F1-score	0.4683	0.4907	0.4653
TPR	0.5969	0.5953	0.5938
FPR	0.1007	0.1011	0.1015
Accuracy Drop	0.3367	0.3382	0.3398
Transferability Score	0.9981	0.8912	0.6948

These results suggest that single-step attacks like FGSM, despite their simplicity, can still cause severe disruption when the surrogate model is well-aligned with the target decision boundary. The higher transferability observed for FGSM implies that the perturbations it generates lie in directions that are highly compatible with the target model's vulnerability regions. Conversely, the reduced transferability of PGD and BIM, which employ iterative refinements, indicates that these methods may generate perturbations that overly exploit the surrogate's specific loss landscape, reducing their generalization to the target. This observation highlights an important trade-off between attack complexity and cross-model effectiveness. It also underscores the need for adaptive adversarial strategies that balance perturbation strength with transferability when attacking non-differentiable models through surrogate-based approaches.

VI. CONCLUSION & FUTURE WORK

In this study, we explored the efficacy of surrogate-based adversarial attacks as a means to bridge the gap between traditional black-box and white-box adversarial strategies. Our results demonstrate that leveraging surrogate models significantly enhances attack success rates, transferability, and computational efficiency compared to conventional black-box attacks. The proposed methodology allows the deployment of gradient-based attacks in black-box settings, revealing previously unseen vulnerabilities in machine learning models, particularly decision tree-based classifiers. These findings emphasize the necessity for more robust adversarial defenses, as many machine learning architectures remain highly susceptible to such attacks. Future research will focus on expanding the scope of surrogate-based adversarial attacks by incorporating a broader range of surrogate models, including transformer-based architectures and more complex deep learning structures. Additionally, we plan to evaluate the effectiveness of these attacks against state-of-the-art adversarial defenses, such as adversarial training, feature squeezing, and certified robustness methods. Another avenue of exploration involves assessing the trade-offs between attack efficiency and detectability to better understand the feasibility of these attacks in real-world adversarial settings. Finally, we aim to refine our methodology by optimizing the surrogate model selection process, ensuring higher attack success rates while minimizing computational overhead.

REFERENCES

[1] D. Yang, Z. Xiao, and W. Yu, "Boosting the adversarial transferability of surrogate models with dark knowledge," in *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2023, pp. 627–635.

[2] Q. Zeng, Z. Wang, Y.-m. Cheung, and M. Jiang, "Ask, attend, attack: A effective decision-based black-box targeted attack for image-to-text models," *arXiv preprint arXiv:2408.08989*, 2024.

[3] T. Wu, T. Luo, and D. C. Wunsch II, "Lrs: Enhancing adversarial transferability through lipschitz regularized surrogate," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 6135–6143.

[4] D. C. Asimopoulos, P. Radoglou-Grammatikis, T. Lagkas, V. Argyriou, I. Moscholios, J. Cani, G. T. Papadopoulos, E. K. Markakis, and P. Sarigiannidis, "Aag: Adversarial attack generator for evaluating the robustness of machine learning models against adversarial attacks," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024, pp. 2682–2689.

[5] D. C. Asimopoulos, P. Radoglou-Grammatikis, I. Makris, V. Mladenov, K. E. Psannis, S. Goudos, and P. Sarigiannidis, "Breaching the defense: Investigating fgsm and ctgan adversarial attacks on iec 60870-5-104 ai-enabled intrusion detection systems," in *Proceedings of the 18th International Conference on Availability, Reliability and Security*, 2023, pp. 1–8.

[6] N. Inkawhich, Y. Xu, A. Mao, Y. Wang, and M. Noseworthy, "Perturbing across the feature hierarchy to improve standard and strict black-box attack transferability," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 4109–4116.

[7] C. Wu, W. Zhang, Y. Zhu, and H. Su, "Skip connections matter: On the transferability of adversarial examples generated with skip connections," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 12 324–12 331.

[8] J. Wang, X. Xia, F. Wei, and J. Xu, "Enhancing the transferability of adversarial attacks through variance tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 192–201.

[9] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9185–9193.

[10] J. Lin, Y. Song, Y. He, and Z. Zhang, "Nesterov accelerated gradient and scale invariance for improving transferability of adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2019.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *International Conference on Learning Representations (ICLR)*, 2015.

[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations (ICLR)*, 2018.

[13] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 39–57.

[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[16] C. Dalamagkas, P. Radoglou-Grammatikis, P. Bouzinis, I. Papadopoulos, T. Lagkas, V. Argyriou, S. Goudos, D. Margounakis, E. Fountoukidis, and P. Sarigiannidis, "Federated detection of open charge point protocol 1.6 cyberattacks," *arXiv preprint arXiv:2502.01569*, 2025.