Enhancing Black-Box Adversarial Attacks on Power System Event Classifiers via Transferability

Yuanbin Cheng, Nanpeng Yu
Department of Electrical and Computer Engineering
University of California, Riverside
Riverside, California 92507 USA
ychen871@ucr.edu, nyu@ece.ucr.edu

Jim Follum
Pacific Northwest National Laboratory
Richland, USA
james.follum@pnnl.gov

***Abstract*—The widespread deployment of phasor measurement units (PMUs) in power transmission systems has accelerated the development of deep learning–based real-time monitoring solutions, such as event classification. Despite these advancements, recent research indicates that adversarial attacks pose a significant threat, as even minor perturbations in the input data can deceive well-trained models. In domains like computer vision, adversarial samples have been shown to transfer across different architectures, thus enabling black-box attacks via surrogate models. However, this transferability phenomenon remains largely unexplored in power system applications. In this work, we conduct a comprehensive study of adversarial transferability for power system event classification using machine learning models and a large-scale dataset. Drawing on the insights from this investigation, we propose a novel ensemble-based black-box adversarial attack that exploits transferability to achieve higher success rates and greater query efficiency. Furthermore, beyond using the Euclidean norm to measure perturbations, we incorporate signal-to-noise ratio (SNR) and maximum mean discrepancy (MMD) to enhance the robustness and depth of our perturbation analysis. Extensive experiments on a large-scale, real-world PMU dataset and state-of-the-art event classifiers highlight the effectiveness of our proposed approach.**

***Index Terms*—Black-box adversarial attack, event classification, phasor measurement units, power system.**

I. Introduction

The rapid deployment of advanced sensors such as phasor measurement units (PMUs) has accelerated the development of data-driven methods in power systems, particularly for event detection and classification [1]. PMUs offer high reporting rates for voltage and current phasors [2]. Their widespread adoption has generated large volumes of data to drive machine learning solutions in power systems [3]. These solutions are crucial to improving the reliability of the modern power grid. Recently, machine learning-based methods have shown remarkable accuracy and efficiency in detecting and classifying transmission grid anomalies, such as voltage, frequency, and oscillation events. Many researchers have developed end-to-end deep neural network-based event classifiers, such as convolutional neural networks (CNNs) [4], enhanced ResNet-50 models [5], and generative adversarial networks (GANs) [6]. Others adopt hierarchical strategies, for example, using a hierarchical CNN with channel filtering [7] or a refined two-level hierarchical CNN-based framework [8].

Despite considerable advances in machine learning-based methods for real-time power systems monitoring, these models remain inherently susceptible to adversarial attacks. They could pose critical reliability and security risks in future deployments. By introducing small perturbations into the input PMU data, adversarial attacks can exploit underlying model weaknesses and trigger incorrect predictions [9]. Recent work by [10] further demonstrates how easily adversarial manipulation can compromise machine learning-based power system event classifiers.

Adversarial attacks are commonly divided into white-box and black-box categories. Both pose significant security threats to machine learning-based models. White-box attacks assume full knowledge of the model and allow for direct gradient computation [9], [11], but real-world constraints often limit this access. In contrast, black-box attacks only require input-output queries and can be subdivided into query-based and transfer-based approaches. Query-based attacks generate adversarial samples by iteratively probing the black-box model using zeroth-order gradient estimation [12], sign gradient estimation [13], [14], or decision boundary analyses [15]–[17]. Transfer-based attacks exploit adversarial transferability [18], [19] by crafting perturbations against a surrogate model, which can then deceive the target. However, their success rates remain limited, particularly when attacking diverse architectures.

While adversarial attacks have been explored in power systems [20], and adversarial transferability has been extensively studied in computer vision, it remains largely underexplored in power system applications, especially in the context of event detection and classification. This paper addresses this gap by conducting a comprehensive study of adversarial transferability for power system event classification models. Drawing on our findings, we propose an ensemble-based black-box attack algorithm that takes advantage of adversarial samples from a surrogate model. This algorithm improves the success rates and query efficiency of existing black-box methods. By identifying how and why machine learning-based models fail under adversarial conditions, our work provides critical insights for developing more robust models and enables the power systems community to strengthen resilience against real-world threats.

Although the Euclidean norm is typically used to quantify

imperceptibility in image-based adversarial attacks, it may not fully capture the impact of perturbations on power system data [21], where even small deviations of the Euclidean norm can produce noticeable noise. To address this limitation, we adopt a multi-metric approach by incorporating two additional measures, signal-to-noise ratio (SNR) and maximum mean discrepancy (MMD), alongside the Euclidean norm. This combination provides a more robust framework for limiting and analyzing adversarial perturbations for power system event detection and classification.

The main contributions of this paper are highlighted below.

- We conduct the first extensive investigation of adversarial transferability in power systems using large-scale datasets and multiple machine learning-based models.
- We propose an ensemble-based method that exploits adversarial transferability to significantly enhance both the success rate and query efficiency of black-box adversarial attacks, surpassing existing techniques.
- We adopt a multi-metric approach, incorporating the Euclidean norm, signal-to-noise ratio (SNR), and maximum mean discrepancy (MMD), to rigorously assess adversarial perturbations, offering deeper insights into their impact on event classification performance.

The remainder of this paper is organized as follows. Section II introduces key notations and formulates the adversarial attack problem. Section III describes our transferability evaluation methodology, proposes an ensemble-based black-box adversarial attack algorithm, and details our multi-metric perturbation analysis approach for power system data. Section IV presents transferability results and compares the performance of our proposed method against several state-of-the-art black-box attack algorithms. Finally, Section V concludes the paper and discusses the directions for future work.

II. Key Notations and Problem Definition

This section begins by defining the key notations used in this paper, including the PMU time series representation, the power system event dataset, and the machine learning-based event classifier. Subsequently, it formalizes the concept of adversarial attacks on classifiers and provides the background of the black-box adversarial attack methodologies.

A. Key Notations

Let $\mathbf{x} = [\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_W]$ represent a time series of PMU measurements spanning a fixed window of length W. Each $\mathbf{m}_i (1 \leq i \leq W)$ is a measurement matrix containing electrical variables (e.g., active power, reactive power, voltage magnitude, and frequency) collected from multiple PMUs.

Each sample $\mathbf{x}$ is paired with an event label y, indicating the type of event captured in the time series. The power system event dataset can be written as:

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots\}.$$

We denote the machine learning-based event classifier by $f_\theta(\cdot)$, where θ represents its learned parameters trained by the above event dataset $\mathcal{D}$. Given an input sample $\mathbf{x}$, the classifier outputs a probability distribution over event types, denoted by $\hat{y} = f_\theta(\mathbf{x})$.

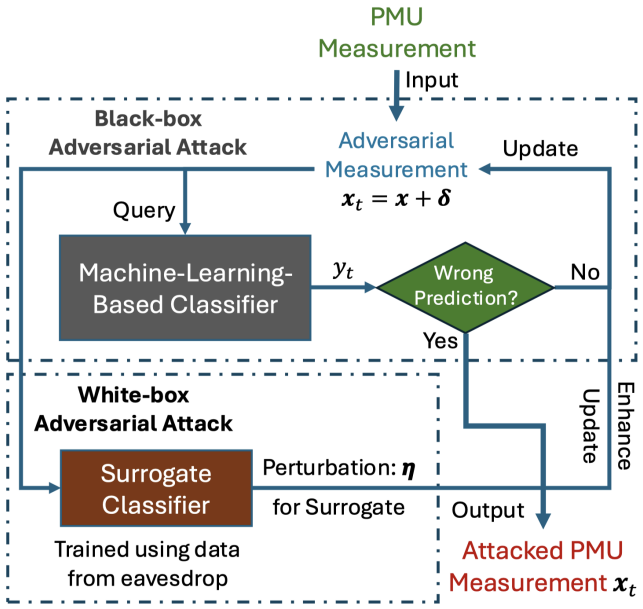

Fig. 1. The illustration of black-box and transfer-based adversarial attacks for PMU measurements.

B. Problem Definition

1) Adversarial Attacks: An adversarial attack on a classifier $f_\theta(\cdot)$ seeks to generate the adversarial sample $\mathbf{x}' = \mathbf{x} + \boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is a small, imperceptible perturbation. Formally, this is expressed as:

$$\arg\max_{\delta} L(f_\theta(\mathbf{x} + \boldsymbol{\delta}), y), \text{ subject to } ||\boldsymbol{\delta}||_2 \leq \epsilon, \qquad (1)$$

where $L(\cdot)$ typically represents the cross-entropy loss, and ϵ constrains the perturbation magnitude to ensure the imperceptibility in terms of the l_2-norm.

C. White-Box vs. Black-Box Attacks

a) White-box attacks: White-box attacks assume full access to the model's architecture and parameters, enabling direct gradient computation. Methods such as FGSM [9] and PGD [11] leverage gradients to generate adversarial perturbations.

b) Black-box attacks: Black-box attacks operate without knowledge of the model's internal details, relying solely on queries and observed outputs (see the upper part of Figure 1). These attacks iteratively refine $\boldsymbol{\delta}$ through trial and error until the classifier misclassifies the perturbed input. Figure 2 illustrates how minor perturbations can deceive a trained power system event classifier.

D. Review of Black-Box Attack Algorithms

a) Score-Based Attacks: These methods exploit confidence scores from the target model to approximate gradients of the loss function $L(\theta; \mathbf{x}, y)$. Attackers craft adversarial samples by iteratively querying the model with small input variations. Notable techniques include Natural Evolutionary Strategies (NES) [22] and ZoSignSGD [12]. Details on score-based attack methods are provided in Section IV-C1.

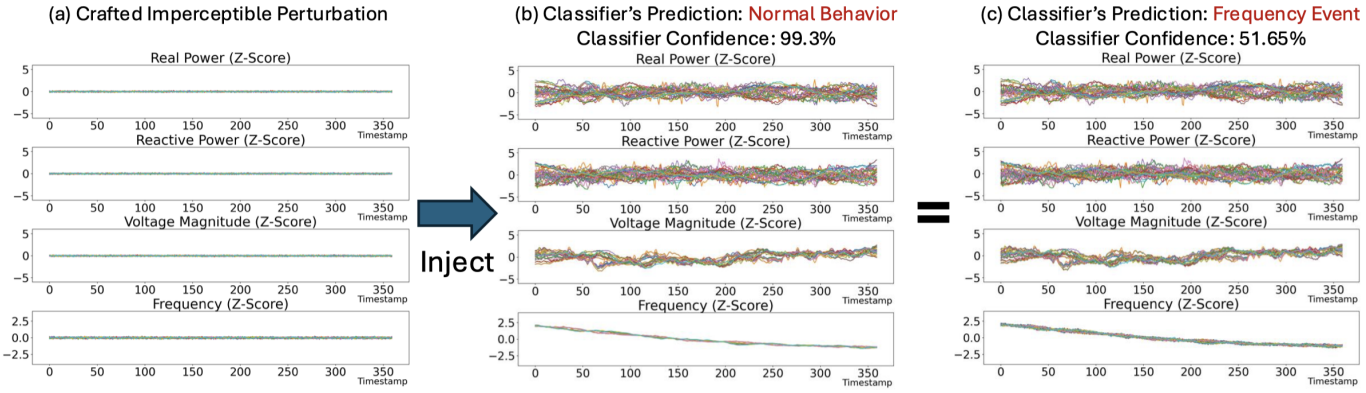

Fig. 2. Example of a successful black-box attack: small perturbation causes misclassification from normal operation to frequency event.

b) Boundary-Based Attacks: These attacks iteratively refine adversarial samples by probing the decision boundary, the hyper-surface separating different classes. By perturbing the input and observing the model output, attackers identify points near this boundary and use them to create nearly identical and misclassified samples [15], [17]. Techniques such as binary search can efficiently locate the boundary points. Boundary-based attacks are presented in detail in Section IV-C1.

III. Technical Method

This section first outlines the evaluation methodology for assessing adversarial transferability between various models. Then, it presents the proposed ensemble-based black-box adversarial attack algorithm, which leverages transferability to amplify adversarial effects in power system event classification. Finally, we introduce two specialized metrics designed specifically for power systems to ensure that the generated perturbations remain imperceptible.

A. Evaluating Adversarial Transferability for Power System Event Classifiers

To assess how well adversarial samples transfer between different event classifiers, we compute the percentage of adversarial samples originally generated to mislead one model that also causes misclassifications in another. We refer to this metric as the *transferability success rate*. A higher rate implies stronger transferability between the two models, reflecting the ability of the same perturbation pattern to exploit shared vulnerabilities across diverse neural architectures.

In our experiments, we measure the transferability success rate across five network architectures: VGG-13 [23], MobileNet-V2 [24], DenseNet-121 [25], ResNet-50, and ResNet-18 [26]. All classifiers are trained on the large-scale dataset described in Section IV-A, which takes approximately 2 hours per model on a Quadro RTX 6000 GPU. Despite architectural differences, all these networks share the same power system event classification objective, enabling a direct comparison of their susceptibility to adversarial samples.

As presented in Section IV-B, our findings reveal that adversarial samples in the power systems domain indeed exhibit transferability across different model architectures. This result underscores the potential of leveraging transferability to strengthen black-box adversarial attacks in power system event classification. Building on these insights, the next subsection introduces a novel ensemble-based black-box adversarial attack algorithm that harnesses transferability to achieve higher success rates and greater query efficiency than state-of-the-art approaches.

B. Ensemble-Based Black-Box Adversarial Attack Algorithm

In this subsection, we first provide an overview of the sign-based black-box attack algorithm [12] and then introduce our proposed ensemble approach that leverages transferability to improve the attack success rate and query efficiency.

Algorithm 1 Sign-based black-box attack

Input: classifier f_θ, data sample $\mathbf{x}$, learning rate α
Output: Adversarial example $\mathbf{x}_{adv}$
Parameters: Perturbation bound ϵ, Maximum iteration N

$\mathbf{x}_{adv} = \mathbf{x}, query\_cnt = 0$
while $query\_cnt < N$ **do**
  $\mathbf{g} = \text{SignGradientEstimate}(f_\theta, \mathbf{x}_{adv})$
  $\mathbf{x}_{adv} = \mathbf{x}_{adv} + \alpha\mathbf{g}$
  if $f_\theta(\mathbf{x}_{adv}) \neq f_\theta(\mathbf{x})$ **then**
    Success, return $\mathbf{x}_{adv}$
  end if
end while
return Fail, return $\mathbf{x}_{adv}$

1) Sign-Based Black-Box Attack: Algorithm 1 illustrates the sign-based attack framework, which iteratively refines adversarial samples by estimating the sign of the loss gradient with respect to the input data. The adversarial example is updated in each iteration by adding a scaled version of this estimated sign. The process continues until the classifier misclassifies the input or the query limit is reached.

Various methods can be used for the gradient estimation. Among these, the state-of-the-art BitSchedule approach [27] offers a reliable and efficient black-box strategy to approximate the sign of the gradient, making it particularly well suited for generating adversarial perturbations.

2) Ensemble-based Black-box Attack via Transferability: Algorithm 2 details our proposed ensemble-based black-box attack strategy, which combines gradient estimates from both

Algorithm 2 Ensemble-based black-box attack algorithm

Input: classifier f_θ, surrogate classifier f_θ^S, data sample $\mathbf{x}$, learning rate α

Output: Adversarial example $\mathbf{x}_{adv}$

Parameters: Perturbation bound ϵ, Maximum iteration N

$\mathbf{x}_{adv} = \mathbf{x}, query\_cnt = 0, iter = 0$
while $query\_cnt < N$ **do**
  $\mathbf{g}_1 = \text{SignGradientEstimate}(f_\theta, \mathbf{x}_{adv})$
  $\mathbf{g}_2 = \text{WhiteBoxAttack}(f_\theta^S, \mathbf{x}_{adv})$
  $\mathbf{g} = \frac{\mathbf{g}_1}{||\mathbf{g}_1||_2} + e^{-iter \cdot d} \frac{\mathbf{g}_2}{||\mathbf{g}_2||_2}$
  $\mathbf{x}_{adv} = \mathbf{x}_{adv} + \alpha \mathbf{g}$
  if $f_\theta(\mathbf{x}_{adv}) \neq f_\theta(\mathbf{x})$ **then**
    Success, stop the attack.
  end if
  $iter$ += 1
end while
return $\mathbf{x}_{adv}$

the black-box model and a white-box surrogate. In each iteration, we compute two gradients:

(1) $\mathbf{g}_1$, the sign of the gradient estimation obtained by querying the black-box target model.

(2) $\mathbf{g}_2$, the gradient estimation derived from a white-box attack on a surrogate model, which is trained using leaked or eavesdropped data by an attacker to approximate the target's behavior due to the transferability of adversarial perturbations.

To combine these two gradients, we introduce an exponentially decaying coefficient $e^{-iter \cdot d}$, designed to give greater weight to the surrogate-based gradient early on and then gradually reduce its influence. Specifically, we form the combined gradient as:

$$\mathbf{g} = \frac{\mathbf{g}_1}{||\mathbf{g}_1||_2} + e^{-iter \cdot d} \frac{\mathbf{g}_2}{||\mathbf{g}_2||_2}, \tag{2}$$

where $iter$ is the current iteration index, and d is a hyperparameter that controls the decay rate (set to 1 in our experiments). In the initial stages of the process, $\mathbf{g}_2$ (from the surrogate model) plays a dominant role, effectively exploiting transferability properties to enhance attack efficacy. As iterations progress, the algorithm is more dependent on $\mathbf{g}_1$, which reflects the direct black-box feedback of the target model.

We can choose from various methods for gradient estimators. In this work, we employ the BitSchedule estimator [27] to approximate the sign of the black-box gradient ($\mathbf{g}_1$), as it provides a reliable and efficient approach to querying the target model. Meanwhile, for the surrogate-based gradient ($\mathbf{g}_2$), we use the Carlini-Wagner L2 attack (C&W) [28], a well-established white-box method known to identify impactful and constrained adversarial perturbations effectively.

C. Perturbation Analysis for PMU Data in Power Systems

Most adversarial attack research evaluates perturbations using the Euclidean norm, assuming that smaller perturbations correspond to more imperceptible changes. However, this assumption does not always hold for power system sensor data [21]. The distinctive patterns in PMU signals make even small Euclidean norm perturbations perceptible, as they introduce noise patterns that can be identified by system operators. To address this limitation, we incorporate two additional metrics for a more comprehensive perturbation analysis: signal-to-noise ratio (SNR) and maximum mean discrepancy (MMD). These metrics provide a more robust assessment of adversarial perturbations, ensuring that attacks remain effective and inconspicuous within real-world power system environments.

1) Signal-to-Noise Ratio (SNR): To estimate the SNR, we adopt the noise filtering procedure from [21], which applies median filters of various orders to the raw PMU measurements. Let $\mathbf{S} = MedianFilter(\mathbf{X})$ represent the filtered signal, and define the noise component as $\mathbf{N} = \mathbf{X} - \mathbf{S}$. The SNR in decibels (dB) is then calculated using the ratio of the standard deviation of the filtered signal $\mathbf{S}$ to the standard deviation of noise $\mathbf{N}$.

$$\text{SNRdB} = 20 \cdot \log_{10} \left(\frac{\text{std}(\mathbf{S})}{\text{std}(\mathbf{N})} \right). \tag{3}$$

Drawing on typical PMU data characteristics, we constrain the SNR of the adversarially perturbed signal to deviate by no more than 1 dB from that of the unperturbed signal.

We chose the 1 dB threshold based on our extensive analysis of the natural variability in the PMU data. Specifically, our two-year study of 40 PMUs, with 100 samples (each containing around 60,000 measurements), revealed that the signal-to-noise ratio (SNR) has a standard deviation of 1.6 dB. This indicates that a 1 dB variation is well within the typical fluctuation range of the system. In other words, a perturbation of up to 1 dB is almost indistinguishable from the inherent measurement noise observed in regular operations.

By enforcing this limit on the adversarial perturbation, we ensure that the noise introduced remains within a standard operating range for power system PMU data, maintaining the integrity and realism of the power system's behavior while still providing a meaningful challenge to the system.

2) Maximum Mean Discrepancy (MMD): MMD is a non-parametric statistical measure that compares two probability distributions using sample data. Let $\mathbf{X}$, $\mathbf{X}'$ be i.i.d. samples from distribution $\mathbf{P}$, and $\mathbf{Y}$, $\mathbf{Y}'$ are i.i.d. samples from distribution $\mathbf{Q}$. Given a positive-definite kernel $k(\cdot, \cdot)$ the squared MMD is defined by:

$$\text{MMD}^2(\mathbf{P}, \mathbf{Q}) = \mathbb{E}[k(\mathbf{X}, \mathbf{X}')] + \mathbb{E}[k(\mathbf{Y}, \mathbf{Y}')] - 2\mathbb{E}[k(\mathbf{X}, \mathbf{Y})]. \tag{4}$$

Intuitively, MMD measures the distance between the mean embeddings of $\mathbf{P}$ and $\mathbf{Q}$ in a reproducing kernel Hilbert space (RKHS). Because it depends only on a suitable kernel and does not require strong assumptions about the underlying data-generating process, MMD is well-suited for detecting both subtle and high-dimensional differences.

In practice, MMD can be coupled with a permutation test to provide a statistically rigorous assessment of whether two datasets come from different distributions. Specifically, one computes the observed MMD in the original samples, then

repeatedly shuffles and partitions a pooled dataset to obtain a distribution of MMD values under the null hypothesis ($\mathbf{P} = \mathbf{Q}$). This yields a p-value indicating how likely the observed MMD would be if $\mathbf{P}$ and $\mathbf{Q}$ are the same.

In the power system context, we apply MMD to ensure that adversarial perturbations do not shift the distribution of PMU signals beyond acceptable limits. Specifically, we require that a hypothesis test comparing the noise distributions before and after the attack yields a p-value above 0.5, indicating that there is insufficient evidence to conclude a significant distributional shift. To make the test highly sensitive to rejecting the null, we set a comparatively large p-value threshold of 0.5. This suggests that the post-attack noise remains statistically indistinguishable from normal variations, minimizing the risk of detection by system operators.

By complementing the Euclidean norm with both SNR and MMD metrics, we gain a more domain-relevant evaluation of adversarial perturbations in PMU data. This multi-metric approach captures both the perceptibility of noise and potential shifts in PMU data distribution.

IV. Numeral Study

In this section, we begin by introducing the large-scale real-world PMU dataset used to train our models. We then present experimental findings on adversarial transferability across different network architectures. Finally, we compare both the success rate and query efficiency of our proposed ensemble-based black-box attack method against seven state-of-the-art baseline algorithms.

A. Dataset and Target ML-based Event Classification Model

This study utilizes two years of PMU data (2016–2017) from the U.S. Western Interconnection, comprising voltage phasors, current phasors, and frequency measurements. 40 PMUs' data are used in this study. Following [6, Section III-F], we cleaned and transformed the raw data into a structured tensor format that captures four variables: active power, reactive power, voltage magnitude, and frequency. The data preprocessing pipeline included discarding unreliable PMUs based on status flags and outlier thresholds, as well as imputing missing data.

Event labels were derived from utility data logs, resulting in 1,204 labeled samples that span four categories: line events, generator events, oscillation events, and normal operation instances. Each sample covers a 12-second interval recorded at 30 Hz, producing a [time steps $\times$ PMUs $\times$ variables] tensor. We trained five neural networks, VGG-13 [23], MobileNetV2 [24], DenseNet-121 [25], ResNet-18, and ResNet-50 [26], on this dataset to perform event classification, serving as target models for the adversarial attacks. The attacks include both white-box evaluations for transferability assessment and black-box attacks.

B. Transferability Result between Different Models

Table I presents the transferability success rates among five different classifiers, revealing that adversarial samples do exhibit some degree of cross-model effectiveness in the power system event classification task. For instance, adversarial samples crafted on VGG13 can still achieve a 34.16% success rate on DenseNet-121 (using DeepFool) and a 20.08% success rate on ResNet-50 (using FGSM). However, these values are substantially lower than those typically observed in image analytics tasks, where transferability between different network architectures can exceed 60%.

A similar pattern emerges in attacks originating from MobileNetV2 and targeting ResNet-50 or ResNet-18, with success rates generally below 30%. This trend underscores the relatively modest cross-architecture transferability observed in power system event classification. One possible reason is the unique data characteristics of PMU signals, which differ significantly from images in both dimensionality and distribution. As a result, adversarial perturbations crafted on one power system event classification model do not consistently align with the decision boundaries of other models.

These findings highlight the inherent difficulty in creating universally transferable adversarial samples for power system event classification, suggesting that techniques proven effective in vision-based domains may not directly translate to this context. Consequently, more specialized or ensemble-based approaches may be required to improve transfer-based attacks on power system event classifiers, as they can better account for the unique features and patterns present in PMU data.

C. Performance of Ensemble-based Black-box Adversarial Attacks

1) Baseline Black-box Attack Algorithms: In this paper, we benchmark six baseline black-box attacks, three score-based and three boundary-based, to compare against our proposed method. SimBA [13] is a straightforward score-based approach that iteratively perturbs individual coordinates and observes the classification output to maximize misclassification. ZoSignSGD [12] also uses score information but applies a zeroth-order optimization strategy, relying on sign-based gradient approximations for efficient parameter updates. Sign-Hunter [14] emphasizes the sign of gradients in the model's loss function, adjusting inputs through a divide-and-conquer approach to effectively mislead the classifier. On the boundary-based side, BoundaryAttack [15] starts with a sample already misclassified and iteratively refines it to minimize the distance to the original data, while OPT Attack [16] employs zeroth-order optimization and binary search to narrow the gap to the decision boundary. Finally, Sign-OPT Attack [17] improves upon OPT by using sign information of distance changes, significantly reducing computational overhead and enhancing query efficiency.

2) Performance Comparison of Ensemble-based Black-box Attacks: Table II and III presents both the success rates (top) and the average query consumption (bottom) for multiple black-box attack algorithms tested against four different classifiers, VGG13, MobileNetV2, DenseNet-121, and ResNet-50, while using ResNet-18 as the surrogate in the ensemble-based

TABLE I
TRANSFERABILITY SUCCESS RATE BETWEEN DIFFERENT MODELS

Surrogate \ Target	Algorithm	VGG13	MobileNetV2	DenseNet-121	ResNet-18	ResNet-50
VGG13	**FGSM**	86.95%	10.97%	17.39%	17.39%	20.08%
	DeepFool	85.92%	24.84%	34.16%	36.64%	26.50%
	C&W	97.10%	9.52%	16.97%	14.28%	16.14%
MobileNetV2	**FGSM**	6.83%	97.51%	8.90%	10.55%	12.83%
	DeepFool	9.31%	99.58%	8.69%	12.42%	12.62%
	C&W	6.83%	97.51%	7.86%	8.28%	8.28%
DenseNet-121	**FGSM**	13.45%	12.83%	95.44%	10.28%	19.25%
	DeepFool	29.81%	26.50%	92.75%	30.43%	22.98%
	C&W	9.73%	11.18%	91.51%	16.35%	17.59%
ResNet-18	**FGSM**	22.77%	39.33%	32.29%	100%	61.28%
	DeepFool	30.64%	49.89%	37.68%	100%	64.38%
	C&W	14.69%	27.32%	23.60%	100%	50.51%
ResNet-50	**FGSM**	13.45%	22.15%	14.69%	40.37%	94.61%
	DeepFool	19.46%	22.36%	18.01%	28.77%	100%
	C&W	10.76%	13.66%	12.42%	24.43%	100%

TABLE II
SUCCESS RATE OF DIFFERENT CLASSIFIERS UNDER – QUERY NUMBER LIMITATION: 1000, PERTURBATION MAGNITUDE LIMITATION: 40, SNR CHANGE THRESHOLD:1dB, MMD $p-value$ THRESHOLD:0.5

	Simba	ZoSignSGD	SignHunter	BoundaryAttack	OPT	Sign-OPT	BitSchedule	Ensemble-based
VGG13	5.17%	11.59%	8.69%	2.48%	1.44%	7.66%	42.65%	**47.01%**
MobileNetV2	13.45%	20.91%	25.25%	7.03%	3.10%	11.18%	57.55%	**64.59%**
DenseNet-121	4.34%	6.83%	13.45%	2.89%	1.03%	2.07%	28.57%	**36.23%**
ResNet-50	6.62%	12.42%	15.73%	2.48%	2.07%	6.00%	42.65%	**63.35%**

TABLE III
AVERAGE QUERY NUMBER OF DIFFERENT CLASSIFIERS UNDER – QUERY NUMBER LIMITATION: 1000, PERTURBATION MAGNITUDE LIMITATION: 40, SNR CHANGE THRESHOLD:1dB, MMD $p-value$ THRESHOLD:0.5

	Simba	ZoSignSGD	SignHunter	BoundaryAttack	OPT	Sign-OPT	BitSchedule	Ensemble-based
VGG13	927	913	892	943	896	891	732	**669**
MobileNetV2	899	877	795	932	945	934	615	**486**
DenseNet-121	928	925	857	934	787	785	790	**677**
ResNet-50	911	900	824	935	801	795	715	**404**

method. Similar results are observed when other classifiers are used as surrogates.

The experiments enforce the query limitation of 1000, the maximum l_2 perturbation norm on the z-scored inputs of 40, the SNR change threshold of 1 dB, and the MMD $p-value$ threshold of 0.5. Under these constraints, the proposed ensemble-based black-box attack (highlighted in red) demonstrates significantly higher success rates than competing methods, ranging from 36.23% on DenseNet-121 to 64.59% on MobileNetV2. This underscores the method's robustness across varying network architectures.

Notably, the ensemble-based attack also excels in query efficiency, requiring fewer queries on average than Simba, ZoSignSGD, SignHunter, BoundaryAttack, OPT, Sign-OPT, and BitSchedule. For example, it uses only 669 queries to achieve higher success rate on VGG13, compared to 732 queries for BitSchedule, and 404 queries on ResNet-50 rather than 795 for Sign-OPT. This reduction in query usage is particularly advantageous in real-world, resource-constrained settings, where there is limited time or computational budgets to perform repeated queries. Combining high success rates with reduced query consumption, the proposed method offers both effectiveness and practical feasibility for adversarial attacks on power systems event classifiers.

D. Progression Query-efficiency Comparison of Ensemble-based Black-box Attacks

Figure 3 illustrates the progression of successful attacks as the query budget increases on ResNet-50, while Tables II and Tables III respectively provide final success rates at the 1000-query limit and average query usage. The figure and above tables show that the proposed ensemble-based black-box attack outperforms all other methods on two key metrics: the total number of successful adversarial samples generated and the efficiency with which those successes are achieved.

In particular, the proposed ensemble-based method reaches a high volume of successful attacks with fewer than 200 queries, surpassing baseline algorithms by a substantial margin even in the early stages. By the 1000-query mark, it maintains its lead, culminating in the highest final success rate while consuming fewer queries per successful attack on average. These

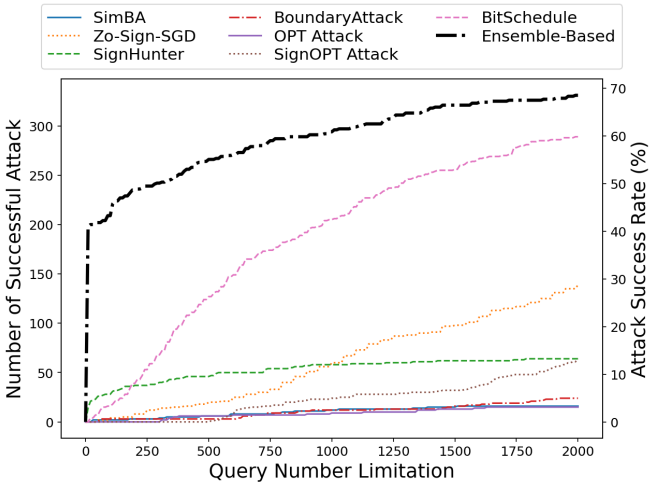

Fig. 3. Number of the successful attack sample with increasing query limitation. (ResNet-50)

results highlight that the integration of transferability through an ensemble-based approach yields substantial benefits when operating under strict query constraints, making it a strong candidate for resource-limited adversarial scenarios.

V. Conclusion

This paper explores adversarial transferability in power system event classification and introduces an ensemble-based black-box attack that improves both success rates and query efficiency. Our findings reveal that adversarial transferability in power systems event classifier is weaker than in vision tasks, necessitating specialized attack strategies. The proposed method effectively combines surrogate gradients with real-time feedback, achieving high success rates and query efficiency across all tested classifiers. Furthermore, by incorporating domain-specific imperceptibility constraints, including SNR and MMD, we ensure that adversarial perturbations remain realistic and undetectable in practical power system operations. Experiments on a large-scale real-world PMU dataset validate the effectiveness of our proposed approach and highlight the significant security implications of adversarial attacks in real-time power system monitoring. Future work will focus on developing more robust defenses and exploring the broader applicability of our technique to other time-series data and mission-critical systems.

References

[1] J. Follum, "Real-time oscillation analysis: Technology readiness, and a vision for future needs and applications," Jun. 2020, Accessed: Sept. 20, 2023. [Online]. Available: https://www.naspi.org/sites/default/files/2020-07/20200624 NASPI Webinar - PJM ESAMS.pdf

[2] A. Monti, C. Muscas, and F. Ponci, *Phasor measurement units and wide area monitoring systems*. Academic Press, 2016.

[3] A. G. Phadke and T. Bi, "Phasor measurement units, WAMS, and their applications in protection and control of power systems," *J. of Mod. Power Syst. and Clean Energy*, vol. 6, no. 4, pp. 619–629, 2018.

[4] W. Wang, H. Yin, C. Chen, A. Till, W. Yao, X. Deng, and Y. Liu, "Frequency disturbance event detection based on synchrophasors and deep learning," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3593–3605, 2020.

[5] J. Shi, B. Foggo, and N. Yu, "Power system event identification based on deep neural network with information loading," *IEEE Trans. Power Syst.*, vol. 36, no. 6, pp. 5622–5632, Nov. 2021.

[6] Y. Cheng, N. Yu, B. Foggo, and K. Yamashita, "Online power system event detection via bidirectional generative adversarial networks," *IEEE Trans. Power Syst.*, vol. 37, no. 6, pp. 4807–4818, 2022.

[7] M. Pavlovski, M. Alqudah, T. Dokic, A. A. Hai, M. Kezunovic, and Z. Obradovic, "Hierarchical convolutional neural networks for event classification on PMU measurements," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[8] L. Zhu, D. J. Hill, and C. Lu, "Hierarchical deep learning machine for power system online transient stability prediction," *IEEE Trans. Power Syst.*, vol. 35, no. 3, pp. 2399–2411, 2020.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd Int. Conf. Learn. Represent., ICLR*, 2015.

[10] Y. Cheng, K. Yamashita, and N. Yu, "Adversarial attacks on deep neural network-based power system event classification models," in *IEEE PES Innovative Smart Grid Technologies - Asia (ISGT Asia)*, 2022, pp. 66–70.

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *6th Int. Conf. Learn. Represent., ICLR*, 2018.

[12] S. Liu, P. Chen, X. Chen, and M. Hong, "signSGD via zeroth-order oracle," in *7th Int. Conf. Learn. Represent., ICLR*, 2019.

[13] C. Guo, J. R. Gardner, Y. You, A. G. Wilson, and K. Q. Weinberger, "Simple black-box adversarial attacks," in *Proc. 36th Int. Conf. Mach. Learn., ICML*, vol. 97, 2019, pp. 2484–2493.

[14] A. Al-Dujaili and U. O'Reilly, "Sign bits are all you need for black-box attacks," in *8th Int. Conf. Learn. Represent., ICLR*, 2020.

[15] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *6th Int. Conf. Learn. Represent., ICLR*, 2018.

[16] M. Cheng, T. Le, P. Chen, H. Zhang, J. Yi, and C. Hsieh, "Query-efficient hard-label black-box attack: An optimization-based approach," in *7th Int. Conf. Learn. Represent., ICLR*, 2019.

[17] M. Cheng, S. Singh, P. H. Chen, P. Chen, S. Liu, and C. Hsieh, "Sign-opt: A query-efficient hard-label adversarial attack," in *8th Int. Conf. Learn. Represent., ICLR*, 2020.

[18] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *5th Int. Conf. Learn. Represent., ICLR*, 2017.

[19] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.

[20] C. Ren, X. Du, Y. Xu, Q. Song, Y. Liu, and R. Tan, "Vulnerability analysis, robustness verification, and mitigation strategy for machine learning-based power system stability assessment model under adversarial examples," *IEEE Trans. Smart Grid*, vol. 13, no. 2, 2022.

[21] M. Brown, M. Biswal, S. Brahma, S. J. Ranade, and H. Cao, "Characterizing and quantifying noise in PMU data," in *IEEE Power and Energy Society General Meeting (PESGM)*, 2016, pp. 1–5.

[22] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *Proc. 35th Int. Conf. Mach. Learn., ICML*, vol. 80, 2018, pp. 2142–2151.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd Int. Conf. Learn. Represent., ICLR*, 2015.

[24] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *IEEE Conf. Comput. Vis. and Pattern Recognit., CVPR*, 2018, pp. 4510–4520.

[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conf. Comput. Vis. and Pattern Recognit., CVPR*, 2017, pp. 2261–2269.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. and Pattern Recognit., CVPR*, 2016, pp. 770–778.

[27] Y. Cheng, K. Yamashita, N. Yu, and Y. Liu, "A hybrid query-efficient black-box adversarial attack on power system event classifiers," in *IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, 2024, pp. 359–365.

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