Black-Box Adversarial Attacks on Smart Grid Stability Score Prediction and a Stochastic Quantile Smoothing Defense

1st Emad Efatinasab
Department of Information Engineering
University of Padua
Padova, Italy
emad.efatinasab@phd.unipd.it

2nd Mirco Rampazzo
Department of Information Engineering
University of Padua
Padova, Italy
mirco.rampazzo@unipd.it

3rd Alessandro Brighente
Department of Mathematics
University of Padua
Padova, Italy
alessandro.brighente@unipd.it

4th Chuadhry Mujeeb Ahmed
Department of Computing
Newcastle University
Newcastle upon tyne, United Kingdom
mujeeb.ahmed@newcastle.ac.uk

5th Mauro Conti
Department of Mathematics
University of Padua, Örebro University
Padova, Italy; Örebro, Sweden
mauro.conti@unipd.it

***Abstract*—Regression-based stability prediction is a core component of decentralized smart grid control, yet its security under adversarial data-injection attacks remains largely unexplored. Unlike prior work that focuses on binary stability classification, many operational systems rely on continuous stability scores that directly influence control decisions. In this paper, we study black-box adversarial attacks against regression-based stability prediction under realistic query and injection budget constraints. We adapt and evaluate multiple query-based derivative-free attacks designed to directly maximize regression output distortion rather than misclassification. We further introduce an uncertainty-aware sampling strategy that allows an attacker to concentrate a limited injection budget on the most vulnerable inputs, significantly amplifying attack impact.**

To mitigate these attacks, we propose Stochastic Quantile Smoothing (SQS), a lightweight, query-side defense that randomizes model outputs using secret quantile sampling and input smoothing. SQS returns a single stochastic but plausible scalar per query, deliberately obstructing gradient and expectation-based black-box optimization while preserving utility for legitimate users. Extensive experiments on smart grid stability prediction show that SQS substantially reduces the effectiveness of black-box attacks by increasing adversarial query cost, while incurring only modest degradation in predictive performance.

***Index Terms*—Adversarial Attacks, Stability Prediction, Smart grids**

I. Introduction

Smart grids represent a major shift in how electricity distribution is managed, bringing new capabilities that modernize traditional power-flow control [1]. To cope with the inherent variability of contemporary power systems, smart grid strategies place strong emphasis on forecasting and other preventative measures that help maintain the balance between supply and demand and reduce the risk of disruptions [2]. One widely adopted mechanism is demand response [3], [4], in particular, Decentralized Smart Grid Control (DSGC) has emerged as a promising approach, as it ties electricity prices to local frequency measurements, allowing price rises and falls to reflect surplus and shortage conditions on the grid [5]. Despite these advances, maintaining stability(in DSGC, the system's ability to restore frequency synchronization and return to a steady operating point after a disturbance) remains challenging: binding price to frequency, heterogeneous participant sensitivities (price elasticities), and varying response/averaging delays complicate real-time control and stability analysis [6]. Grid instability can seriously disrupt electricity provision and the economy, ranging from damaging voltage fluctuations to cascading large-scale blackouts [7]. Stability prediction systems have been developed as critical software tools that analyze operational data to anticipate instability and enable proactive control actions. Machine learning (ML) and AI methods have proven especially effective for this task, with prior studies reporting high predictive accuracy across diverse modeling approaches [8]. Despite these advancements, significant challenges remain. While ML and AI-based methods significantly enhance stability prediction, they also expand the attack surface and are susceptible to adversarial manipulation, an issue that has only recently begun to receive attention in the context of AI-driven smart grid stability prediction [2], [7], [9].

Most existing studies frame stability assessment as a binary classification problem, labeling system states as stable or unstable, but we expect DSGC's continuous stability score to be the primary operational signal for grid controllers, with binary classification serving only as a secondary, high-level alarm rather than the main control input. While useful for high-level monitoring, DSGC relies on continuous stability-related signals that directly drive control actions, with binary labels typically derived from these values. Regression outputs convey

the severity, direction, and temporal evolution of instability and therefore directly inform operational decisions, distinguishing mild deviations that allow gradual correction from severe events that require immediate intervention. Errors in predicted stability scores can delay or misdirect responses and worsen system stress. Despite this high-stakes role, adversarial attacks literature has focused mainly on classification [10], leaving regression-based stability prediction comparatively underexplored, a concerning gap given the potential for manipulation to distort control actions and compromise grid resilience.

Contributions. Against these backdrops, we introduce a realistic, regression-oriented threat model for smart-grid stability prediction in which adversaries directly maximize error on continuous stability scores, and we evaluate this model with adapted black-box attacks parameterized by an injection magnitude (per-sample perturbation bound) and an injection budget (number of inputs modified), with every perturbation clipped to physically plausible sensor ranges. To make attacks practical under tight budgets, we propose an uncertainty-aware sample-selection strategy that ranks inputs by a fused score of data density (outlier detection) and local prediction sensitivity, and show that targeting the top-B samples yields substantially larger error increases per perturbed sample than random selection. For defense, we train predictors with simultaneous quantile regression [11] to obtain calibrated intervals and deploy *Stochastic Quantile Smoothing (SQS)* at inference: SQS returns one stochastic but plausible scalar per query by combining input smoothing, secret quantile sampling, and randomized convex mixing of symmetric quantile outputs. SQS is a lightweight, API-level defense (no retraining, ensembles, or model internals required) that preserves utility for benign users while dramatically increasing attacker query complexity and obfuscating gradient information. Code and data are available at:

https://github.com/emadef1/Black-Box-Adversarial-Attacks

II. Related Works

Advanced AI techniques now play a central role in stability analysis and control for smart grids [12]. For example, Breviglieri et al. [13] show that optimized deep models can forecast stability by analyzing the DSGC system across a range of input conditions. A Multidirectional LSTM architecture for stability forecasting is proposed by [8]. Convolutional neural networks have also been applied; for example, [14] employs a CNN for transient-stability assessment and for predicting instability modes. Finally, whereas much of the literature frames the problem as binary classification, only a few works (e.g. [15], [16]) treat stability prediction as a regression task. Relying solely on binary labels (stable/unstable) may limit the amount of actionable information available to grid operators, as it does not convey how close the system is to instability. Recent work consistently shows that ML methods used in power systems are vulnerable to adversarial manipulation [17]. Aygül et al. [18] report substantial drops in performance during cyber attacks, with transient-stability predictors suffering sharp accuracy declines under attack. Song et al. [19] provide a detailed study of adversarial examples in voltage-stability assessment for the New England 10-machine test system. Despite these findings, Most adversarial ML work targets classification, leaving regression problems, such as predicting continuous stability scores, relatively neglected [10]. This gap matters because manipulating a continuous score can subtly change control decisions or pricing and slowly erode system resilience in ways that are more difficult to detect.

III. System Model

We ground our study in the Decentralized Smart Grid Control (DSGC) paradigm and describe the concrete regression task we address, predicting a continuous stability score that quantifies how close the network is to losing synchronous equilibrium. In DSGC, electricity prices are coupled to grid frequency and reflect the instantaneous balance between supply and demand; stability corresponds to the system's ability to return to a synchronous operating point after a disturbance [2]. The publicly available DSGC dataset [20] includes a continuous stability metric (stab), defined as the maximum eigenvalue of the system's linearized dynamics, alongside a derived binary label (stabf) indicating stable versus unstable conditions. Prior work has shown that the continuous stability score is a valid, physically meaningful learning target, using both regression and classification formulations on the same data [15], [16]. Accordingly, we model stability prediction as a supervised regression task in which a predictor maps compact grid-state and control features to a single real-valued stability score, trained offline on historical nominal data. This formulation captures not only whether the system is unsafe, but when and by how much it drifts toward unsafe operating regimes, which is critical for downstream monitoring and decision-making.

IV. Threat Model

We consider adversarial data injection attacks against a regression-based stability score prediction system deployed in a DSGC setting. The attacker's objective is to degrade prediction accuracy, either by increasing overall error or by inducing systematically misleading stability assessments that influence operational decisions. Such attacks can cause the system to overestimate instability, triggering unnecessary protective actions (e.g., reserve activation, load shedding, or emergency control), leading to operational disruption, increased costs, and degraded service quality. Conversely, underestimating instability may prevent timely intervention, increasing the risk of sustained oscillations, loss of synchronization, or cascading outages. We assume an adversary can inject falsified measurement samples into the smart grid data stream. Such false data injection attacks can be carried out in several ways, for instance via man-in-the-middle compromises that intercept and alter communications between field devices and control systems [21], or by exploiting vulnerabilities in IEC 61850 Ethernet communications within distribution substations [22]. Following established practice in the smart-grid adversarial robustness literature, we explicitly constrain the adversary's capabilities, as unconstrained attacks or perturbations affecting

all samples are not considered realistic threat models [2]. We assume the attacker can manipulate only a limited fraction of samples (e.g., at most 4,000 out of 12,000). This fraction is sufficient to induce measurable performance degradation while remaining covert and unlikely to trigger existing monitoring or anomaly detection mechanisms. Second, the perturbation applied to each manipulated sample is bounded, with a maximum magnitude of 0.2 per feature. Relative to the empirical feature ranges in the dataset, this corresponds to changes between $\approx 3.51\%$ and $\approx 5.77\%$ of a feature's full observed range, ensuring that injected data remain small, physically plausible, and statistically consistent with normal operating conditions. Moreover, all attack injections are explicitly bounded and clipped to the empirical sensor ranges.

We consider a **black-box** attacker, as access to model internals (the white-box assumption) is comparatively unlikely in real-world [2]. In this scenario, the adversary has no access to the model architecture, parameters, or internal representations and can only query the deployed predictor and observe the returned continuous stability score. The attacker, therefore, must rely on query-based optimization techniques to construct effective perturbations. This attack could happen by realistic intrusion paths against grid control systems (e.g., specialised ICS malware that establishes remote interaction with control infrastructure such as CRASHOVERRIDE [23], and compromising a grid control center via ICS-specific malware such as Industroyer [24] which was used in the Ukrainian power grid attack).

V. Methodology

A. Stability Prediction Model

Prior work has used data-driven regression models like feed-forward neural networks [15], [16] to predict grid stability indices and to support online control decisions. To be aligned with the literature, the stability predictor used in this work is a single feed-forward neural regression model that maps a compact feature vector $x \in \mathbb{R}^d$ to a scalar stability score $\hat{s}$. The model serves two purposes at inference time: (i) produce a point estimate of the stability score and (ii) provides calibrated prediction intervals, which we leverage to implement a defense policy against black-box query adversarial attacks. The predictor is a feed-forward neural regressor whose input is the original feature vector $x \in \mathbb{R}^d$ augmented with a scalar quantile-conditioning variable $\tau \in [0, 1]$, i.e. $\tilde{x} = [x,\ (\tau - 0.5) \cdot 12] \in \mathbb{R}^{d+1}$. The network uses six fully connected ReLU layers (256, 512, 768, 1024, 1280 neurons) and a final linear readout producing the stability score $\hat{s}$.

We train the network with Simultaneous Quantile Regression (SQR) in [11]. Instead of training a separate model per quantile, SQR trains a single model $f_\theta(x, \tau)$ that returns the τ-th conditional quantile of the target for any $\tau \in [0, 1]$. SQR minimizes the expected pinball (quantile) loss:

$$\mathcal{L}_{\text{SQR}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\tau \sim \mathcal{U}[0,1]} \big[\ell_\tau \big(f_\theta(x_i, \tau), y_i \big) \big],$$

We train using the Adam optimizer, a batch size of 256, and 700 epochs. Simultaneous quantile regression captures heteroskedastic and non-Gaussian observation noise without resorting to ensembles, and provides well-calibrated prediction intervals in practice [11].

B. Adversarial Attacks

As described in Section IV, we evaluate black-box attacks (derivative-free optimizers) adapted from classification to a regression setting. Since regression models do not expose decision boundaries, all attacks are reformulated to maximize the magnitude of the output displacement rather than classification error. Specifically, given an input x, all attacks optimize either

$$\big| f(x + \delta, \tau) - f(x, \tau) \big| \quad \text{or} \quad \big(f(x + \delta, \tau) - f(x, \tau) \big)^2,$$

subject to a norm constraint $\|\delta\|_2 \leq \epsilon$ and feature-wise box constraints.

- **NES (Natural Evolution Strategies).** NES [25] is a population-based gradient-estimation method that samples random Gaussian perturbations around an input, evaluates the model on these perturbed points, and constructs a gradient estimate by correlating the perturbations with observed changes in the objective. Unlike classification variants that optimize confidence margins or loss values, we apply NES to directly maximize the squared change in the regression output, $\max_{\|\delta\|_2 \leq \epsilon} \big(f(x+\delta, \tau) - f(x, \tau) \big)^2$. Each update is normalized, followed by an ℓ_2 projection and feature-wise clipping to maintain validity. We use a population size of 50, a noise scale of $\sigma = 0.01$, and run the attack for 30 iterations.
- **Square Attack.** The Square Attack [26] is a randomized black-box method that explores structured perturbations. Originally developed for image classification, where it iteratively modifies random square regions and greedily accepts proposals that increase the attack objective, we adapt the strategy to vector and tabular inputs by randomly selecting a subset of features each iteration and applying uniform perturbations to those coordinates. Each candidate perturbation is projected onto the ℓ_2 ball and clipped to valid feature ranges before evaluation. The attack objective is the squared regression output displacement. We run the attack for 30 iterations with a block size of 50.
- **SimBA (Simple Black-box Attack).** The SimBA [27] attack is a lightweight, coordinate-wise method that performs greedy one-dimensional searches along randomly chosen input directions. Originally designed to reduce the probability of the correct class by stepping along basis vectors, our regression adaptation iteratively selects a single feature and tests positive and negative perturbations of fixed magnitude, accepting the update if it increases the squared change in the regression output. After each accepted step, the perturbation is projected onto the ℓ_2 constraint set and clipped to valid feature bounds with a fixed step size of 0.01 and 30 iterations.

- **ZOO (Zeroth-Order Optimization).** The Zeroth-Order Optimization (ZOO) attack [28] estimates gradients via finite differences along selected input dimensions. At each iteration, ZOO perturbs a subset of coordinates, measures the resulting change in the regression objective, and constructs a sparse gradient estimate. For regression, we replace the classification loss with the squared output displacement and estimate gradients around the current perturbed input. The gradient estimate is normalized and used for a projected update under the ℓ_2 constraint, followed by feature-wise clipping. Gradients are estimated using 50 coordinates per iteration with step size 0.01 over 30 iterations.

C. Uncertainty-Aware Sampling

We assume a fixed attack budget of at most 4,000 modified samples. A naive black-box attacker would choose these samples uniformly at random, but this overlooks the fact that some inputs are inherently more vulnerable to perturbations than others. By identifying such "weak" inputs, an attacker can use its limited budget more effectively. To this end, we introduce a principled sample selection strategy that ranks test inputs without access to model internals or epistemic/aleatoric uncertainty, and prioritizes those most likely to yield high attack impact.

a) Density-based score: We detect out-of-distribution (low-density) samples using an Isolation Forest trained on the attacker-visible test inputs (data going through the grid network). Let the clean test set be denoted by $\mathcal{D}_{\text{test}} = \{x_i\}_{i=1}^{N}$. For each input x_i we define a raw density-based vulnerability score $d(x_i) = -\,\mathrm{IF}(x_i)$ where $\mathrm{IF}(\cdot)$ denotes the Isolation Forest decision function. By construction, larger values of $d(x_i)$ indicate that x_i lies in a low-density (or outlier) region of the input space and is therefore more atypical relative to the bulk of the test distribution. This low-density signal serves as a proxy for epistemic uncertainty. When the model has seen few (or no) training examples in a neighbourhood around x_i, its predictions are less reliable, and its behaviour is more difficult to anticipate. Intuitively, such blind-spot inputs often correspond to regions of the input space where the learned function extrapolates poorly, so a small adversarial perturbation can more readily produce a large change in the model output. Isolation Forests are a practical choice for constructing this score in the attacker-visible setting as they are unsupervised, scale well to moderate-dimensional data, and explicitly measure how isolated a point is without requiring labels.

b) Local sensitivity score: We additionally quantify the local smoothness of the model in a neighbourhood around each input. Intuitively, if the model output $f(x)$ varies rapidly under small input perturbations, then even tiny adversarial changes can induce large prediction errors. To estimate this effect, we sample M small Gaussian perturbations around each test input x_i, $x_i^{(j)} = x_i + \epsilon_j$, with $\epsilon_j \sim \mathcal{N}(0, \sigma^2 I)$ and evaluate the model outputs $f(x_i^{(j)}, \tau)$. The local sensitivity score is defined as the empirical standard deviation of these predictions,

$$
\begin{aligned}
s_{\text{sens}}(x_i) &= \sqrt{\frac{1}{M}\sum_{j=1}^{M}\big(f(x_i+\epsilon_j,\tau) - \bar{f}_i\big)^2}, \\
\bar{f}_i &= \frac{1}{M}\sum_{j=1}^{M} f(x_i+\epsilon_j,\tau).
\end{aligned}
\tag{1}
$$

This quantity serves as an empirical estimate of the local Lipschitz behaviour of f at x_i: larger values of $s_{\text{sens}}(x_i)$ indicate that the model output changes substantially under small random perturbations, suggesting a steeper local gradient. From a robustness perspective, a small local gradient (or Lipschitz constant) implies stability if:$|f(x+\delta) - f(x)| \le L\|\delta\|$. Then an adversary must apply a sufficiently large perturbation δ to induce a significant change in the output.

Both scores are min–max normalized across the test set and the final attacker-visible vulnerability score is defined as

$$
s(x_i) = w_{\text{dens}}\,\tilde{d}(x_i) + w_{\text{sens}}\,\tilde{s}(x_i), \tag{2}
$$

This score $s(x)$ ranks inputs by overall vulnerability. The weights $w_{\text{dens}}, w_{\text{sens}}$ (0.5, 3.0 based on hyperparameter optimization) control the trade-off between outlierness and gradient sensitivity.

D. Stochastic Quantile Smoothing (SQS) for Adversarial Defense

Many regression APIs expose a single fixed output (e.g., the median or mean). This deterministic oracle makes it easy for black-box attackers to estimate gradients via finite differences or NES, as repeated queries yield nearly identical values, so an attacker quickly approximates directional derivatives. Stochastic Quantile Smoothing (SQS) disrupts this by returning one random draw per query instead of a fixed point estimate. SQS is composed of the following three ingredients.

1) Input smoothing (Gaussian noise) : Each query input x is perturbed by a small random Gaussian noise $\epsilon_x \sim \mathcal{N}(0, \sigma^2 I)$ and the model is evaluated at $x + \epsilon_x$. Even a tiny noise injection can dramatically confuse finite-difference and population estimators. In practice, the noise adds a roughly constant variance term to any finite-difference estimate

$$
\frac{f(x+h) - f(x)}{h}, \tag{3}
$$

So the estimated derivative has high variance. Here, h denotes the small, attacker-chosen finite-difference step.

2) Secret quantile sampling : Instead of always returning the same quantile (e.g., the median at $\tau = 0.5$), SQS randomly selects a quantile τ from a hidden pool on each query. For input $x + \epsilon_x$, we compute the lower and upper quantiles

$$
q_\ell = q(x+\epsilon_x, \tau), \qquad q_h = q(x+\epsilon_x, 1-\tau). \tag{4}
$$

The actual value of τ (and hence which quantile function is used) is secret. To the attacker, the model appears to be a mixture of functions rather than a fixed map. This non-stationarity means an attacker cannot assume a single deterministic output function; they would have to infer the hidden τ-distribution or average over many queries to cancel out the randomness.

3) Random convex mixing: Finally, SQS returns a convex combination of the two symmetric quantiles:

$$\tilde{f}(x) = \alpha q_\ell + (1-\alpha) q_h, \qquad \alpha \sim \text{Uniform}(0,1). \quad (5)$$

This guarantees plausibility of the output: because q_ℓ and q_h bracket the model's predictive interval, any convex mix lies between them. In practice, if $\alpha = 0.5$, the mix is exactly the median. The random mixing injects additional variance into $\tilde{f}(x)$ but keeps the result within the model's own quantile range, avoiding out-of-range or obviously incorrect values.

A key goal is to maintain output plausibility and avoid large bias. A naive defense could simply add random noise to the model's scalar output. However, arbitrary additive noise risks producing impossible values (outside the predictive interval) or large bias in expectation. SQS's strategy is different: by sampling within the model's own quantile range, the random outputs remain calibrated. The expected SQS response satisfies

$$\mathbb{E}[\tilde{f}(x)] = \frac{1}{2}\big(q(x,\tau) + q(x,1-\tau)\big), \quad (6)$$

which lies between the two quantiles. By contrast, adding, say, Gaussian noise to the output could easily produce values outside the model's plausible range. Quantile mixing thus preserves plausibility as every returned value is inside the model's predictive interval. We tune the noise scale and quantile pool per model to trade utility for robustness; using a small $\sigma = 0.05$ and the spread-out pool (0.01,0.025,0.05,0.1,0.25,0.5,0.75,0.9,0.95,0.975) yields only a modest increase in prediction error while greatly increasing the number of queries an attacker must make.

VI. Evaluation

Let $\{(x_i, y_i)\}_{i=1}^n$ be the test set, where $y_i \in \mathbb{R}$ is the ground-truth target and $\hat{y}_i$ denotes the model prediction (either on clean or adversarial inputs). We evaluate regression performance using Mean Absolute Error (MAE), the coefficient of determination R^2, and Normalized MAE (NMAE).

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|,$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$\text{NMAE} = \frac{\text{MAE}}{\frac{1}{n}\sum_{i=1}^n |y_i|} \times 100\%,$$

$$\Delta_{\text{MAE}} = \frac{\text{MAE}_{\text{adv}} - \text{MAE}_{\text{clean}}}{\text{MAE}_{\text{clean}}} \times 100\%.$$

TABLE I
Adversarial evaluation results (4,000 injected samples per attack, $\epsilon = 0.2$).

Attack	MAE	NMAE	R^2	ΔMAE
Clean	0.0020	6.01%	0.9873	0.0%
NES	0.0046	13.84%	0.9595	130.36%
Square	0.0028	8.35%	0.9832	38.92%
SimBA	0.0024	7.40%	0.9856	23.09%
ZOO	0.0046	13.93%	0.9588	131.85%

A. Dataset

The experiments use an augmented version of the Electrical Grid Stability Simulated Dataset from the UCI Machine Learning Repository [20], a widely used benchmark for stability analysis and adversarial studies in power systems [2], [29]. The dataset contains 60,000 simulated instances generated from a 4-node star network under the DSGC paradigm. Each sample includes 12 primary predictive variables and two dependent variables that capture stability-related outcomes. We split the data into training (80%), and test (20%) sets, and we standardize all input features using a training-set StandardScaler.

B. Stability Prediction Model Evaluation

In the non-adversarial setting and before the introduction of SQS defense, we report results using the median (0.5-quantile) prediction, which provides a robust point estimate that is less sensitive to outliers than the mean. Under this setting, our model achieves an MAE of 0.0020, an NMAE of 6.01%, and a coefficient of determination of $R^2 = 0.9873$, indicating high predictive accuracy and strong agreement between predicted and ground-truth stability values.

C. Attacks Performance Evaluation

We assess adversarial impact using a fixed perturbation budget $\epsilon = 0.2$ by randomly injecting 4,000 adversarial samples per attack into the test set and evaluating performance over the full dataset. This setup models a realistic threat where only a small fraction of inputs can be manipulated to degrade overall system behavior, rather than a worst-case evaluation limited to adversarial samples. As shown in Table I, all black-box attacks cause noticeable degradation compared to the clean baseline, with substantially varying severity. NES and ZOO are the most damaging attacks. NES increases the MAE to 0.0046 (NMAE $= 13.84\%$), corresponding to a relative MAE increase of $+130.36\%$ and a reduction in R^2 to 0.9595. ZOO yields a comparable level of degradation, with MAE $= 0.0046$ (NMAE $= 13.93\%$), resulting in a $+131.85\%$ increase in MAE and a drop in R^2 to 0.9588. By contrast, structure-aware and coordinate-based attacks induce more moderate performance degradation.

The observed differences reflect each attack's optimization strategy. Gradient-estimating methods such as NES and ZOO generate distributed perturbations that better exploit a smooth regression loss, producing substantially larger MAE increases and sharper drops in R^2. By contrast, Square and SimBA rely

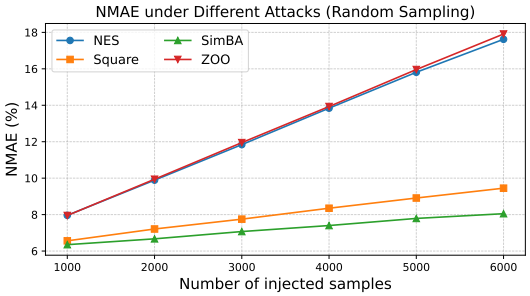

Fig. 1. NMAE of the stability regression model under different adversarial attacks as the number of injected adversarial samples(random sample selection) increases from 1000 to 6,000.

on structured block updates or single-coordinate greedy steps, which are less effective at maximizing continuous prediction error averaged over a large test set. Because metrics are aggregated across the entire test set, these results show that modest, targeted sample injections (e.g., ZOO increases MAE to approximately 2.3× the clean baseline) can significantly reduce performance, a serious concern for safety-critical systems that rely on numeric stability scores.

We further perform an ablation by varying the adversary's injection budget from 1,000 to 6,000 samples in 1,000-sample increments while keeping all attack hyperparameters fixed; the resulting NMAE trends (Fig. 1) show a consistent, monotonic drop in regression accuracy across all attack types, confirming that even partial data injection is sufficient to significantly degrade stability prediction. Gradient-estimation black-box attacks are consistently the most damaging. NES increases NMAE from 7.96% at 1,000 injected samples to 17.62% at 6,000 samples, corresponding to a nearly threefold increase relative to the clean baseline (6.01%). ZOO exhibits a nearly identical trajectory, rising from 7.95% to 17.91% over the same range and producing the largest overall degradation. In contrast, structure-aware and coordinate-based attacks have a noticeably weaker effect. Square increases NMAE more gradually, from 6.56% at 1,000 samples to 9.45% at 6,000 samples, while SimBA remains the least harmful, reaching only 8.05% NMAE at the maximum injection budget.

To illustrate stealthiness and impact, Fig. 2 visualizes random sensor features before and after manipulation; the perturbations remain small and within operational bounds (Sec. IV), so manipulated samples are difficult to distinguish from benign data. We evaluated standard unsupervised anomaly detectors on raw inputs, Isolation Forest, One-Class SVM, and Gaussian Mixture models, and found they perform poorly against these attacks, yielding near-random ROC AUCs (for example, Isolation Forest achieves TPR = 0.197 and AUC = 0.523 on NES), which underscores the need for alternative defenses.

D. Uncertainty-Aware Sampling Evaluation

In this section, we evaluate the performance of the attacks when combined with our uncertainty-aware sampling strategy, replacing random sample selection. We analyze the effect of

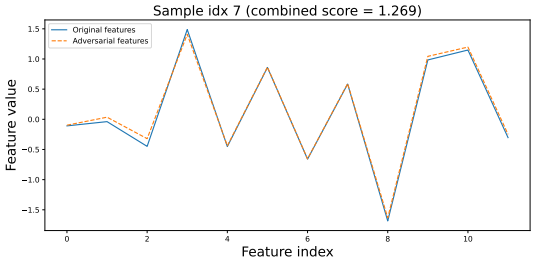

Fig. 2. Example sensor feature trajectories before and after adversarial manipulation by ZOO

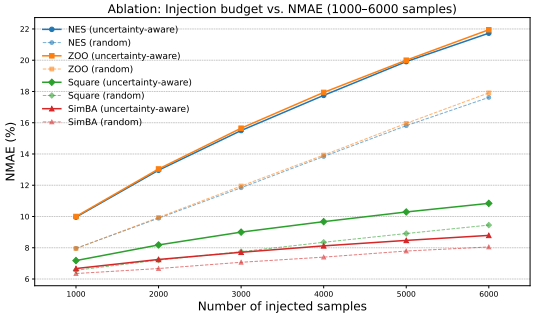

Fig. 3. Comparing the black-box attacks under random sampling and uncertainty-aware sampling strategies. The plot reports the NMAE of the stability predictor as the adversarial sample injection budget increases.

varying the adversarial injection budget from 1000 to 6,000 samples in increments of 1000, and report the corresponding results in Fig. 3. Across all four attacks, uncertainty-aware sampling consistently amplifies the damage caused by adversarial injections compared to random sampling. For NES, the NMAE increase attributable to uncertainty-aware selection ranges from approximately 2.0 percentage points at a budget of 1,000 samples to over 4.1 percentage points at 6,000 samples. A similar trend is observed for ZOO, where the gap between uncertainty-aware and random sampling widens steadily as the injection budget grows. These results indicate that the benefit of informed sample selection compounds as the adversary gains the ability to corrupt more inputs. The effect is present but more modest for structure-aware attacks such as Square and SimBA. While both attacks already induce smaller absolute errors under random sampling, uncertainty-aware selection still yields consistent improvements across all budgets, increasing NMAE by roughly 0.6–1.4 percentage points depending on the budget. This suggests that even attacks with constrained perturbation patterns benefit from prioritizing inputs that are more influential under the model's predictive uncertainty. Under our primary threat model with a fixed query budget of 4,000 samples, uncertainty-aware NES increases the model's NMAE from 13.84% (random sampling baseline) to 17.75%, while ZOO increases NMAE from 13.93% to 17.94%. Importantly, these gains are achieved without modifying the attack optimization and solely from changing which samples are selected for injection.

TABLE II
ADVERSARIAL EVALUATION ON THE *attacked subset* (2000 INJECTED SAMPLES PER ATTACK, $\varepsilon = 0.2$).

	Deterministic (median)			SQS (stochastic)		
Attack	**NMAE**	$\mathbf{R^2}$	**ΔMAE**	**NMAE**	$\mathbf{R^2}$	**ΔMAE**
Clean	9.70%	0.9741	N/A	20.22%	0.9404	N/A
NES	74.99%	0.5316	672.92%	35.94%	0.8546	77.72%
Square	30.08%	0.9084	210.00%	20.58%	0.9392	1.74%
SimBA	21.37%	0.9464	120.26%	20.22%	0.9405	0.00%
ZOO	75.79%	0.5223	681.17%	21.44%	0.9336	6.00%

E. Stochastic Quantile Smoothing (SQS) Evaluation

In this section, We evaluate SQS against uncertainty-aware black-box attacks using an adapted, apples-to-apples protocol that compares attacker success under a deterministic median oracle versus the SQS oracle. First, we measure the cost of SQS for benign queries on the full test set to establish an acceptable utility baseline. With the deterministic median ($\tau = 0.5$), the predictor attains an MAE of 0.0020 (NMAE $= 6.01\%$) and an R^2 score of 0.9873. Under SQS, MAE increases to 0.0032 (NMAE $= 9.64\%$) while the R^2 score remains high at 0.9813. Thus, SQS introduces a controlled increase in error and variance while preserving strong explanatory power and calibration prior to measuring attack resilience.

Because each query involves stochastic input perturbations, secret quantile sampling, and randomized output mixing, large-scale black-box attacks become substantially more expensive. Consequently, we evaluate the top 2,000 uncertainty-selected vulnerable points (rather than 4,000) and compute all metrics on this attacked subset by comparing model outputs on the same inputs before and after the attack. This protocol isolates the per-sample effect produced by the attack and prevents conflating injection sparsity with global test-set averages. To ensure fair comparisons, we preserve the attack hyperparameters and optimizer settings used earlier. We evaluate two attacker-knowledge regimes: (1) a *deterministic oracle* (baseline), in which the adversary queries the fixed median quantile ($\tau = 0.5$) and constructs perturbations from deterministic responses, representing a low-variance API; and (2) an *SQS oracle* (deployed), in which the adversary interacts with the production API and receives a single stochastic draw per query, induced by input noise, secret τ sampling, and random convex mixing. Crucially, the adversary does not know the defender's secret quantile pool, the per-query τ realizations, or the per-call mixing coefficients.

Because Table II reports metrics only on the uncertainty-selected attacked subset (2000 samples), the results isolate per-sample attack and defense effects. Under the deterministic median oracle, gradient-estimation and finite-difference attacks produce the largest local damage, NES increases NMAE by +65.29 percentage points (a relative rise of +672.92% in MAE) and ZOO by +66.09 pp (+681.17%). Structured and coordinate-wise attacks are less destructive as Square increases MAE by +20.38 pp (+210.00%) and SimBA by +11.67 pp (+120.26%). Stochastic Quantile Smoothing (SQS) substantially reduces attack success. Under SQS NES's absolute increase falls to +15.72 pp (+77.72% relative), ZOO to +1.22 pp (+6.0%), Square to +0.36 pp (+1.8%), and SimBA to essentially 0.00 pp (0.0%). SQS trades a small decrease in clean explanatory power (clean R^2: $0.9741 \rightarrow 0.9404$) for a large preservation of explanatory power under attack. Under the deterministic median oracle, NES and ZOO reduce attacked-subset R^2 to 0.5316 and 0.5223, respectively, whereas SQS raises these to 0.8546 (+0.3230) and 0.9336 (+0.4113). For Square and SimBA, the effect is smaller: Square's R^2 improves modestly ($0.9084 \rightarrow 0.9392$, +0.0308), while SimBA remains essentially unchanged. These results show that SQS not only reduces absolute adversarial error but also keeps attacked outputs substantially more correlated with true targets, whereas a low-variance deterministic oracle can render attacked outputs largely uninformative.

Empirically, SQS sharply reduces the success of single-evaluation attacks like ZOO, SimBA, and Square because their core acceptance or derivative tests become noisy under per-query randomness. NES, which estimates directions by averaging across perturbation populations, is more robust but can only restore performance by increasing population size or iterations. In practice, SQS transforms most black-box attacks from cheap to query-expensive, NES remains the most effective adversary but demands substantially larger budgets and can become impractical under realistic rate limits and monitoring, and its localized impact can be further obscured when performance is averaged over the full test set rather than evaluated on the attacked subset. SQS is not without cost. On the attacked subset, the clean (non-adversarial) NMAE increases from 9.70% under the deterministic median oracle to 20.22% under SQS. The observed increase is larger on the attacked subset because it contains inputs that are intrinsically more vulnerable and were deliberately selected by the attacker using uncertainty-aware criteria. Importantly, the resulting utility loss is a controllable policy choice: parameters such as the input-noise scale σ, the τ-pool, and the mixing strategy can be tuned to trade off a smaller benign accuracy loss for a correspondingly smaller reduction in attacker effectiveness.

a) Computational Cost: A single SQS query remains linear in the input dimension: it performs two quantile evaluations plus $O(D)$ overhead for input noise and mixing, so a legitimate request costs $O(D)$ (a modest constant factor above an undefended forward pass). Black-box attacks, however, must issue many queries to estimate gradients or search the input space: NES and ZOO scale as $O(P\,T\,D)$ and $O(K\,T\,D)$, respectively, where P is the NES population size, K is the number of coordinates sampled per ZOO iteration, and T is the number of iterations. Structured heuristics such as Square or SimBA scale as $O(T\,D)$. Thus, SQS preserves $O(D)$ defender cost per request but turns a single evaluation into $O(q \cdot D)$ attacker work (with $q = PT$ for NES or $q = KT$ for ZOO), imposing a multiplicative query-complexity burden on attackers even though per-query complexity remains unchanged.

b) Complementarity Considerations: Finally, SQS is a lightweight, black-box defense that operates at the query inter-

face and requires no retraining, architectural changes, or access to model internals. It is designed to complement, not replace, other defenses (e.g., query detection, rate limiting), allowing seamless combination with stronger system-level guarantees.

VII. Conclusion

We study black-box adversarial attacks on regression-based smart-grid stability prediction and show that small, stealthy perturbations, selected by an uncertainty-aware sample selection strategy that prioritizes low-density, high-sensitivity inputs, can induce operationally meaningful prediction errors even under strict injection/query budgets. Gradient-estimating attacks (e.g., NES, ZOO) are especially effective, causing substantial degradation in global performance metrics. To mitigate this, we propose Stochastic Quantile Smoothing (SQS): an API-level defense that uses randomized input smoothing, secret quantile sampling, and convex mixing within predictive intervals to return a single stochastic yet plausible output per query. SQS raises attacker variance and non-stationarity (increasing query complexity) while preserving model utility, and it can be deployed directly on existing quantile-regression models without retraining or ensembles.

VIII. Acknowledgment

The project is funded in equal parts by the University of Padua and VIMAR S.p.A. under project *AISec4IoT*.

References

[1] R. Bayindir, I. Colak, G. Fulli, and K. Demirtas, "Smart grid technologies and applications," *Renewable and sustainable energy reviews*, vol. 66, pp. 499–516, 2016.

[2] E. Efatinasab, N. Azadi, G. A. Susto, C. Mujeeb Ahmed, and M. Rampazzo, "Fortifying smart grid stability: Defending against adversarial attacks and measurement anomalies," *Sustainable Energy, Grids and Networks*, vol. 43, p. 101799, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235246772500181X

[3] P. Palensky and D. Dietrich, "Demand side management: Demand response, intelligent energy systems, and smart loads," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 3, pp. 381–388, 2011.

[4] M. B. Gorzałczany, J. Piekoszewski, and F. Rudziński, "A modern data-mining approach based on genetically optimized fuzzy systems for interpretable and accurate smart-grid stability prediction," *Energies*, vol. 13, no. 10, 2020.

[5] B. Schäfer, M. Matthiae, M. Timme, and D. Witthaut, "Decentral smart grid control," *New journal of physics*, vol. 17, no. 1, p. 015002, 2015.

[6] V. Arzamasov, K. Böhm, and P. Jochem, "Towards concise models of grid stability," in *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, 2018, pp. 1–6.

[7] E. Efatinasab, A. Brighente, M. Rampazzo, N. Azadi, and M. Conti, "Gan-grid: A novel generative attack on smart grid stability prediction," in *Computer Security – ESORICS 2024*, J. Garcia-Alfaro, R. Kozik, M. Choraś, and S. Katsikas, Eds. Cham: Springer Nature Switzerland, 2024, pp. 374–393.

[8] M. Alazab, S. Khan, S. S. R. Krishnan, Q.-V. Pham, M. P. K. Reddy, and T. R. Gadekallu, "A multidirectional lstm model for predicting the stability of a smart grid," *IEEE Access*, vol. 8, pp. 85 454–85 463, 2020.

[9] Q. Song, R. Tan, C. Ren, Y. Xu, Y. Lou, J. Wang, and H. B. Gooi, "On credibility of adversarial examples against learning-based grid voltage stability assessment," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 2, pp. 585–599, 2024.

[10] X. Kong and Z. Ge, "Adversarial attacks on regression systems via gradient optimization," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 12, pp. 7827–7839, 2023.

[11] N. Tagasovska and D. Lopez-Paz, "Single-model uncertainties for deep learning," *Advances in neural information processing systems*, vol. 32, 2019.

[12] Z. Shi, W. Yao, Z. Li, L. Zeng, Y. Zhao, R. Zhang, Y. Tang, and J. Wen, "Artificial intelligence techniques for stability analysis and control in smart grids: Methodologies, applications, challenges and future directions," *Applied Energy*, vol. 278, p. 115733, 2020.

[13] P. Breviglieri, T. Erdem, and S. Eken, "Predicting smart grid stability with optimized deep models," *SN Computer Science*, vol. 2, pp. 1–12, 2021.

[14] Z. Shi, W. Yao, L. Zeng, J. Wen, J. Fang, X. Ai, and J. Wen, "Convolutional neural network-based power system transient stability assessment and instability mode prediction," *Applied Energy*, vol. 263, p. 114586, 2020.

[15] Z. Allal, H. N. Noura, O. Salman, and K. Chahine, "Leveraging the power of machine learning and data balancing techniques to evaluate stability in smart grids," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108304, 2024.

[16] F. Ucar, "A comprehensive analysis of smart grid stability prediction along with explainable artificial intelligence," *Symmetry*, vol. 15, no. 2, 2023.

[17] Y. Chen, Y. Tan, and D. Deka, "Is machine learning in power systems vulnerable?" in *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, 2018, pp. 1–6.

[18] K. Aygul, M. Mohammadpourfard, M. Kesici, F. Kucuktezcan, and I. Genc, "Benchmark of machine learning algorithms on transient stability prediction in renewable rich power grids under cyber-attacks," *Internet of Things*, vol. 25, p. 101012, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2542660523003359

[19] Q. Song, R. Tan, C. Ren, and Y. Xu, "Understanding credibility of adversarial examples against smart grid: A case study for voltage stability assessment," in *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*, ser. e-Energy '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 95–106.

[20] V. Arzamasov, "Electrical Grid Stability Simulated Data ," UCI Machine Learning Repository, 2018, DOI: https://doi.org/10.24432/C5PG66.

[21] O. Sen, D. van der Velde, P. Linnartz, I. Hacker, M. Henze, M. Andres, and A. Ulbig, "Investigating man-in-the-middle-based false data injection in a smart grid laboratory environment," in *2021 IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, 2021, pp. 01–06.

[22] R. Macwan, C. Drew, P. Panumpabi, A. Valdes, N. Vaidya, P. Sauer, and D. Ishchenko, "Collaborative defense against data injection attack in iec61850 based smart substations," in *2016 IEEE Power and Energy Society General Meeting (PESGM)*, 2016, pp. 1–5.

[23] Cybersecurity and Infrastructure Security Agency (CISA), "Crashoverride malware," 2017, accessed: 2025-03-12. [Online]. Available: https://www.cisa.gov/news-events/alerts/2017/06/12/crashoverride-malware

[24] L. Salazar, S. R. Castro, J. Lozano, K. Koneru, E. Zambon, B. Huang, R. Baldick, M. Krotofil, A. Rojas, and A. A. Cardenas, "A tale of two industroyers: It was the season of darkness," in *2024 IEEE Symposium on Security and Privacy (SP)*, 2024, pp. 312–330.

[25] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *International conference on machine learning*. PMLR, 2018, pp. 2137–2146.

[26] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: a query-efficient black-box adversarial attack via random search," in *European conference on computer vision*. Springer, 2020, pp. 484–501.

[27] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, "Simple black-box adversarial attacks," in *International conference on machine learning*. PMLR, 2019, pp. 2484–2493.

[28] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 15–26.

[29] J. Mulo, P. Tian, A. Hussaini, H. Liang, and W. Yu, "Towards an adversarial machine learning framework in cyber-physical systems," in *2023 IEEE/ACIS 21st International Conference on Software Engineering Research, Management and Applications (SERA)*. IEEE, 2023, pp. 138–143.